

Discriminant analysis with errors in variables

Sébastien Loustau* and Clément Marteau†

Abstract

The effect of measurement error in discriminant analysis is investigated. Given observations $Z = X + \epsilon$, where ϵ denotes a random noise, the goal is to predict the density of X among two possible candidates f and g . We suppose that we have at our disposal two learning samples. The aim is to approach the best possible decision rule G^* defined as a minimizer of the Bayes risk.

In the free-noise case ($\epsilon = 0$), minimax fast rates of convergence are well-known under the margin assumption in discriminant analysis (see [24]) or in the more general classification framework (see [30, 2]). In this paper we intend to establish similar results in the noisy case, i.e. when dealing with errors in variables. In particular, we discuss two possible complexity assumptions that can be set on the problem, which may alternatively concern the regularity of $f - g$ or the boundary of G^* . We prove minimax lower bounds for these both problems and explain how can these rates be attained, using in particular Empirical Risk Minimizer (ERM) methods based on deconvolution kernel estimators.

1 Introduction

In the problem of discriminant analysis, we usually observe two i.i.d. samples $X_1^{(1)}, \dots, X_n^{(1)}$ and $X_1^{(2)}, \dots, X_m^{(2)}$. Each observation $X_j^{(i)} \in \mathbb{R}^d$ is assumed to admit a density with respect to a σ -finite measure Q , dominated by the Lebesgue measure. This density will be denoted by f if the observation belongs to the first set (i.e. when $i = 1$) or g in the other case. Our aim is to infer the density of a new incoming observation X . This problem can be considered as a particular case of the more general and extensively studied binary classification problem (see [12] for a detailed introduction or [7] for a concise survey).

In this framework, a decision rule or classifier can be identified with a set $G \subset \mathbb{R}^d$, which attributes X to f if $X \in G$ and to g otherwise. Then, we can associate to each classifier G its corresponding Bayes risk $R_K(G)$ defined as

$$R_K(G) = \frac{1}{2} \left[\int_{K/G} f(x) dQ(x) + \int_G g(x) dQ(x) \right], \quad (1.1)$$

where we restrict the problem to a compact set $K \subset \mathbb{R}^d$. The minimizer of the Bayes risk (the best possible classifier for this criterion) is given by

$$G_K^* = \{x \in K : f(x) \geq g(x)\}, \quad (1.2)$$

where the infimum is taken over all subsets of K . The Bayes classifier is obviously unknown since it explicitly depends on the couple (f, g) . The goal is thus to estimate G_K^* thanks to a classifier $\hat{G}_{n,m}$ based on the two learning samples.

In this paper we propose to estimate the Bayes classifier G_K^* defined in (1.2) when dealing with noisy samples. For all $i \in \{1, 2\}$, we assume that we observe

$$Z_j^{(i)} = X_j^{(i)} + \epsilon_j^{(i)}, \quad j = 1, \dots, n_i, \quad (1.3)$$

*Université d'Angers, LAREMA, loustau@math.univ-angers.fr

†INSA de Toulouse, IMT, marteau@math.univ-toulouse.fr

instead of the $X_j^{(i)}$. The $\epsilon_j^{(i)}$ denotes random variables expressing measurement errors. We will see in this paper that we are faced to an inverse problem, and more precisely to a deconvolution problem. Indeed, if ϵ admit a density η with respect to the Lebesgue measure, then the corresponding density of the $Z_j^{(i)}$ is the convolution product $(f.\mu) * \eta$ if $i = 1$ or $(g.\mu) * \eta$ if $i = 2$, provide that $dQ(x) = \mu(x)dx$ for some bounded function μ . It gives rise to a deconvolution step in the estimation procedure. Deconvolution problems arise in many fields where data are obtained with measurements errors and are at the core of several nonparametric statistical studies. For a general review of the possible methodologies associated to these problems we may mention for instance [26]. More specifically, we refer to [14] in density estimation or [8] where goodness-of-fit tests are constructed in the presence of noise. The main key of all these studies is to construct a deconvolution kernel which may allow to annihilate the noise ϵ . More details on the construction of such objects are provided in Section 3. It is important to note that in this discriminant analysis setup, or more generally in classification, there is up to our knowledge no such a work. The aim of this paper is to describe minimax rates of convergence in noisy discriminant analysis under the margin assumption.

In the free-noise case, i.e. when $\epsilon = 0$, [24] has attracted the attention on minimax fast rates of convergence (i.e. faster than $n^{-\frac{1}{2}}$) and states in particular

$$\inf_{\hat{G}} \sup_{G_K^* \in \mathcal{G}(\alpha, \rho)} \left[R_K(\hat{G}) - R_K(G_K^*) \right] \approx n^{-\frac{\alpha+1}{2+\alpha+\rho\alpha}}, \text{ as } n \rightarrow +\infty, \quad (1.4)$$

where $\mathcal{G}(\alpha, \rho)$ is a non parametric set of candidates G_K^* with complexity $\rho > 0$ and margin parameter $\alpha \geq 0$ (see Section 2.1 for a precise definition). In (1.4), the complexity parameter $\rho > 0$ is related to the notion of entropy with bracketing whereas the margin is used to relate the variance to the expectation. It allows [24] to get improved bounds using the so-called peeling technique of [16]. This result is at the origin of a recent and vast litterature of fast rates of convergence in classification (see for instance [25, 2]) or in general statistical learning (see [19]). In these papers, the complexity assumption can be of two forms: geometric assumption over the class of candidates G_K^* (such as finite VC dimension, or boundary fragments) or assumptions on the regularity of the regression function of classification (plug-in type assumptions). In [25], minimax fast rates are stated for finite VC class of candidates whereas plug-in type assumptions have been studied in classification in [2] (see also [12, 28]). More generally [19] proposes to consider $\rho > 0$ as a complexity parameter in local Rademacher complexities and gives general upper bounds generalizing (1.4) and the results of [24] and [2].

In all these results, empirical risk minimizers appear as good candidates to reach these fast rates of convergence. Indeed, given a class of candidates \mathcal{G} , a natural way to estimate G_K^* is to consider an Empirical Risk Minimization (ERM) approach. In standard discriminant analysis (e.g. in the free-noise case considered in [24]), the risk $R_K(G)$ in (1.2) can be estimated by

$$R_{n,m}(G) = \frac{1}{2n} \sum_{i=1}^n \mathbf{1}_{\{X_i^{(1)} \in G^C\}} + \frac{1}{2m} \sum_{i=1}^m \mathbf{1}_{\{X_i^{(2)} \in G\}}, \quad (1.5)$$

leading to an empirical risk minimizer $\hat{G}_{n,m}$, if it exists, defined as:

$$\hat{G}_{n,m} = \arg \min_{G \in \mathcal{G}} R_{n,m}(G). \quad (1.6)$$

Unfortunately, in the error-in-variable model, since we observe noisy samples $Z = X + \epsilon$, the probability densities of the observed variables w.r.t. the Lebesgue measure are respectively convolution $(f.\mu) * \eta$ and $(g.\mu) * \eta$. As a result, classical ERM principle fails since:

$$\frac{1}{2n} \sum_{i=1}^n \mathbf{1}_{\{Z_i^{(1)} \in G^C\}} + \frac{1}{2m} \sum_{i=1}^m \mathbf{1}_{\{Z_i^{(2)} \in G\}} \longrightarrow \frac{1}{2} \left[\int_{K/G} (f.\mu) * \eta(x) dx + \int_G (g.\mu) * \eta(x) dx \right] \neq R_K(G).$$

As a consequence, we propose to add a deconvolution step in the classical ERM procedure by considering the solution of the minimization:

$$\min_{G \in \mathcal{G}} R_{n,m}^\lambda(G),$$

where $R_{n,m}^\lambda(G)$ is an asymptotically unbiased estimator of $R_K(G)$ which uses kernel deconvolution estimators with smoothing parameter λ . It is called deconvolution empirical risk and will be of the form

$$R_{n,m}^\lambda(G) = \frac{1}{2n} \sum_{j=1}^n h_{G^C, \lambda}(Z_j^{(1)}) + \frac{1}{2m} \sum_{j=1}^m h_{G, \lambda}(Z_j^{(2)}), \quad (1.7)$$

where the $h_{G, \lambda}(\cdot)$ are smoothed versions of indicator functions used in classical ERM for direct observations (see Section 3 for details).

In this paper, we would like to describe as precisely as possible the influence of the error ϵ on the classification rates of convergence and the presence of fast rates. Our aim is to use the asymptotic theory of empirical processes in the spirit of [16] (see also [32]) when dealing with the deconvolution empirical risk (1.7). To this end, we give the explicit form of functions $h_{G, \lambda}$ in this framework. In particular, we need to study in details the complexity of the class of functions $\{h_{G, \lambda}, G \in \mathcal{G}\}$ in order to get statistical performances of the solution of the ERM estimator. This complexity is related to the imposed complexity over \mathcal{G} , such as boundary fragment assumptions or regularity hypothesis on the function $f - g$. For each assumption, we establish lower and upper bounds and discuss the performances of this deconvolution ERM estimator for this problem. Such a study allows a first comparison of the robustness of these complexity assumptions w.r.t. the presence of errors in variables. Remark that the results presented here focus on the discriminant analysis set up but could be generalized to the classification framework in a future work. Moreover the problem of adaptation will not be considered in this paper but could be the core of a more advanced contribution.

We point out that the definition of the empirical risk (1.7) leads to a new and interesting theory of risk bounds detailed in Section 3 for discriminant analysis. In particular, the parameter λ has to be calibrated to reach a bias/variance trade-off in the decomposition of the excess risk. Related ideas have been recently proposed in [18] in the gaussian white noise model and density estimation setting for more general linear inverse problem using singular value decomposition. In our framework, up to our knowledge, the only minimax result is [17] which gives minimax rates in Hausdorff distance for manifold estimation in the presence of noisy variables. [11] gives also consistency and limiting distribution for estimators of boundaries in deconvolution problems, but no minimax results are proposed. In the direct case, we can also apply this methodology and consider an empirical risk given by the estimation of f and g using simple kernel density estimators. This idea has been already mentioned in [33] in the general learning context and called Vicinal Risk Minimization (see also [9]). However even in pattern recognition and in the direct case, up to our knowledge, there is no asymptotic rates of convergence for this empirical minimization principle.

The paper is organized as follows. In Section 2, the two main complexity assumptions used in this paper are explicated and associated lower bounds are proposed. These lower bounds generalize the previous lower bounds of [24] and [2]. Deconvolving ERM attaining these rates are presented in Section 3. A brief discussion and some perspectives are gathered in Section 4 while Section 5 is dedicated to the proofs of the main results.

2 Plugin vs boundary fragments

In this section, we detail some common assumptions (complexity and margin) that can be set on the pair (f, g) . We then propose lower bounds on the corresponding minimax rates.

First of all, given a set $G \subset K$, simple algebra indicates that the excess risk $R_K(G) - R_K(G^*)$ can be written as:

$$R_K(G) - R_K(G_K^*) = \frac{1}{2} d_{f,g}(G, G_K^*),$$

where the pseudo-distance $d_{f,g}$ over subsets of $K \subset \mathbb{R}^d$ is defined as

$$d_{f,g}(G_1, G_2) = \int_{G_1 \Delta G_2} |f - g| dQ,$$

and $G_1 \Delta G_2 = [G_1^c \cap G_2] \cup [G_2^c \cap G_1]$ is the symmetric difference between two sets G_1 and G_2 . In this context, there is another natural way of measuring the accuracy of a decision rule G through the quantity:

$$d_\Delta(G, G_K^*) = \int_{G \Delta G_K^*} dQ,$$

where d_Δ defines also a pseudo-distance on the subsets of $K \subset \mathbb{R}^d$.

In this paper, we are interested in the minimax rates associated to these pseudo-distances. In other words, given a class \mathcal{F} , one would like to quantify as precisely as possible the corresponding minimax risks defined as

$$\inf_{\hat{G}_{n,m}} \sup_{(f,g) \in \mathcal{F}} d_\square(\hat{G}_{n,m}, G_K^*),$$

where the infimum is taken over all possible estimators of G_K^* and d_\square stands for $d_{f,g}$ or d_Δ following the context. In particular, we will exhibit classification rules $\hat{G}_{n,m}$ attaining these rates. In order to obtain a satisfying study of the minimax rates mentioned above, one need to detail the considered classes \mathcal{F} . Such a class expresses some conditions that can be set on the pair (f, g) . They are often separated in two categories: margin and complexity assumptions.

A first condition that can be set on the pair (f, g) is the well-known margin assumption. It has been introduced in discriminant analysis (see [24]) as follows:

Margin Assumption: *There exists positive constants $t_0, c_2, \alpha \geq 0$ such that for $0 < t < t_0$:*

$$Q\{x \in K : |f(x) - g(x)| \leq t\} \leq c_2 t^\alpha. \quad (2.1)$$

This assumption is related to the behaviour of $|f - g|$ at the boundary of G_K^* . It may give a variety of minimax fast rates of convergence which depends on the margin parameter α . A large margin corresponds to configurations where the slope of $|f - g|$ is high at the boundary of G_K^* . The most favorable case corresponds to a margin $\alpha = +\infty$ when $f - g$ jumps at the boundary of G_K^* .

From a practical point of view, this assumption provides a precise description of the interaction between the pseudo distance $d_{f,g}$ and d_Δ . In particular, it allows a control of the variance of the empirical processes involved in the upper bounds. Note that in the presence of noise in variables, Lemma 4 in Appendix proposes an usefull generalization of Lemma 2 in [24]. More general assumptions of this type can be formulated (see for instance [6] or [19]) in a more general statistical learning context.

The margin assumption is 'structural' in the sense that it describes the difficulty to distinguish an observation having density f from an other with density g . In order to provide a complete study, one also needs to set an assumption on the difficulty to find G_K^* in a possible set of candidates, namely a complexity assumption. In the classification framework, two different kind of complexity assumptions are often proposed in the literature. The first kind concerns the regularity of the boundary of the Bayes classifier. Indeed, our aim is to estimate G_K^* , which yet corresponds to a nonparametric set estimation problem. In this context, it seems natural to traduce the difficulty of the learning process by condition on the shape of G_K^* . An other way to describe the complexity of the problem is to impose condition on the regularity of the underlying densities f and g . Such kind of condition is originally related to plug-in approaches.

Remark that any clear connexion can be established between such kind of assumption: a set G_K^* with a smooth boundary is not necessarily associated to smooth densities. In the two following subsections, we provide a precise description of the assumptions that we will use in this paper. In each case, we propose lower bounds for the associated minimax rates of convergence in this noisy setting. Corresponding upper bounds are presented and discussed in Section 3.

For the sake of convenience, we will also require in the following an additional assumption on the noise ϵ . We assume in the sequel that $\epsilon = (\epsilon_1, \dots, \epsilon_d)'$ admit a density η with respect to the Lebesgue measure satisfying

$$\eta(x) = \prod_{i=1}^d \eta_i(x_i) \quad \forall x \in \mathbb{R}^d. \quad (2.2)$$

In other word, the entries of the vector ϵ are independent. The assumption below describes the difficulty of the considered problems. It is often called the ordinary smooth case in the inverse problem litterature.

Noise Assumption: *There exist $(\beta_1, \dots, \beta_d)' \in \mathbb{R}_+^d$ such that for all $i \in \{1, \dots, d\}$, $\beta_i > 1/2$,*

$$|\mathcal{F}[\eta_i](t)| \sim |t|^{-\beta_i}, \text{ and } |\mathcal{F}'[\eta_i](t)| \sim |t|^{-\beta_i} \text{ as } t \rightarrow +\infty,$$

where $\mathcal{F}[\eta_i]$ denotes the Fourier transform of the η_i . Moreover, we assume that $\mathcal{F}[\eta_i](t) \neq 0$ for all $t \in \mathbb{R}$ and $i \in \{1, \dots, d\}$.

Classical results in deconvolution (see e.g. [14], [15] or [8] among others) are stated for $d = 1$. Two different settings are then distinguished concerning the difficulty of the problem which is expressed through the shape of $\mathcal{F}[\eta]$. One considers alternatively the case where $|\mathcal{F}[\eta](t)| \sim |t|^{-\beta}$ as $t \rightarrow +\infty$, which yet corresponds to mildly ill-posed inverse problem or $|\mathcal{F}[\eta](t)| \sim e^{-\gamma t}$ as $t \rightarrow +\infty$ which leads to a severely ill-posed inverse problem. This last setting corresponds to a particularly difficult problem and is often associated to low minimax rates of convergence.

In this paper, we only deal with d -dimensional mildly ill-posed deconvolution problems. For the sake of brevity, we do not consider severely ill-posed inverse problems or possible intermediates (e.g. a combination of polynomial and exponential decreasing densities). Nevertheless, the rates in these cases could be obtained through the same steps.

2.1 The boundary fragment assumption

We focus in this subsection on an assumption related to the regularity of the boundary of G_K^* . More precisely, we deal with the family of boundary fragments on $K = [0, 1]^d$. A set $G \subset [0, 1]^d$ belongs to a class of boundary fragments (see [20]) if there exists $b : [0, 1]^{d-1} \rightarrow [0, 1]$ such that:

$$G = \{x = (x_1, \dots, x_d) : x_d \leq b(x_1, \dots, x_{d-1})\} := G_b.$$

For given $\gamma, L > 0$ the class of Hölder boundary fragments is then defined as

$$\mathcal{G}(\gamma, L) = \{G_b, b \in \Sigma(\gamma, L)\},$$

where $\Sigma(\gamma, L)$ is the class of isotropic Hölder continuous functions $b(x_1, \dots, x_{d-1})$ having continuous partial derivatives up to order $\lfloor \gamma \rfloor$, the maximal integer strictly less than γ and such that:

$$|b(y) - p_{b,x}(y)| \leq L|x - y|^\gamma, \forall x, y \in [0, 1]^{d-1},$$

where $p_{b,x}$ is the Taylor polynomial of b at order $\lfloor \gamma \rfloor$ at point x .

Boundary fragment assumption. *There exist γ_f and L positive constants such that the set G_K^* belongs to $\mathcal{G}(\gamma_f, L)$.*

In the following, we denote by $\mathcal{F}_{\text{frag}}$ the set of all pairs (f, g) satisfying both the *margin* and *boundary fragment* assumptions. Theorem 1 states lower bounds for the minimax risks over the class $\mathcal{F}_{\text{frag}}$. The proof is postponed to Section 5.

Theorem 1 *Let $K = [0, 1]^d$ and $\mathcal{F} = \mathcal{F}_{\text{frag}}$. Suppose that Q is the Lebesgue measure on K and that the noise assumption is satisfied. Then*

$$\liminf_{n \rightarrow +\infty} \inf_{\hat{G}_{n,m}} \sup_{(f,g) \in \mathcal{F}_{\text{frag}}} (n \wedge m)^{\tau_d(\alpha, \beta, \gamma_f)} d_{\square}(\hat{G}_{n,m}, G_K^*) > 0,$$

where the infimum is taken over all possible estimators of the set G_K^* and

$$\tau_d(\alpha, \beta, \gamma_f) = \begin{cases} \frac{\gamma \alpha}{\gamma_f(2 + \alpha) + (d-1)\alpha + 2\alpha \sum_{i=1}^{d-1} \beta_i + 2\alpha \beta_d \gamma_f} & \text{for } d_{\square} = d_{\Delta} \\ \frac{\gamma(\alpha + 1)}{\gamma_f(2 + \alpha) + (d-1)\alpha + 2\alpha \sum_{i=1}^{d-1} \beta_i + 2\alpha \beta_d \gamma_f} & \text{for } d_{\square} = d_{f,g}. \end{cases}$$

Remark that we obtain exactly the same lower bounds as [24] in the direct case, which yet corresponds to the situation where $\beta_j = 0$ for all $j \in \{1, \dots, d\}$. In this particular framework, the minimax rate of convergence mainly depends on γ_f and α . The coefficient γ_f corresponds to the regularity of the boundary of G_K^* . Greater is γ_f , easier is the estimation. The term α is related to the margin assumption. The case $\alpha = +\infty$ actually corresponds to a jump of the function $f - g$ near the boundary of G_K^* . On the opposite hand, a small α is associated to a very difficult problem since the difference between f and g may be quite small in such a situation.

In the presence of noise in the variables, the rates obtained in Theorem 1 are slower. The price to pay is an additional term of the form

$$2\alpha \sum_{i=1}^{d-1} \beta_i + 2\alpha \beta_d \gamma_f.$$

This term clearly connects the difficulty of the problem to the values of the coefficients β_1, \dots, β_d . Moreover the above expression highlights a connection between the margin parameter and the ill-posedness. The role of the margin parameter over the inverse problem can be summarized as follows. Higher is the margin, higher is the price to pay for a given degree of ill-posedness. When the margin parameter is small, the problem is difficult at the boundary of G_K^* and we can only expect a non-sharp estimation of G_K^* . In this case it is not significantly worst to add noise.

On the contrary, for large margin parameter, there is nice hope to give a sharp estimation of G_K^* and then perturb the inputs variables have strong consequences in the performances.

Remark also in the above expression that the first $d - 1$ components of ϵ have not the same impact as the last (vertical) component. This is due to the fact that we consider boundary fragments with a given regularity γ_f .

2.2 The plug-in assumption

The boundary fragment assumption concerns the set G_K^* and in particular the smoothness of its boundary. Other conditions have been proposed in the literature in order to explain and quantify the difficulty related to a classification problem.

An alternative hypothesis concerns the regularity of the function $f - g$ itself. In the following, we denote by $\Sigma'(\gamma, L)$ the class of d -dimensional isotropic Hölder continuous functions.

Plug-in Assumption. *There exists γ_p and L' positive constants such that $f - g \in \Sigma'(\gamma_p, L')$.*

We then call $\mathcal{F}_{\text{plug}}$ the set of all pairs (f, g) satisfying both the *margin* and *plug-in* assumptions, since the previous assumption is often associated to plug-in rules in the statistical learning literature. The following theorem proposes a lower bound for the noisy smooth discriminant analysis problem in such a setting.

Theorem 2 *Let $\mathcal{F} = \mathcal{F}_{\text{plug}}$. Suppose that Q is absolutely continuous with respect to the Lebesgue measure and that the noise assumption is satisfied. Then, provided $\alpha \leq 1$,*

$$\liminf_{n \rightarrow +\infty} \inf_{\hat{G}_{n,m}} \sup_{(f,g) \in \mathcal{F}_{\text{plug}}} (n \wedge m)^{\tau'_d(\alpha, \beta, \gamma)} d_{\square}(\hat{G}_{n,m}, G_K^*) > 0,$$

where the infimum is taken over all possible estimators of the set G_K^* and

$$\tau'_d(\alpha, \beta, \gamma) = \begin{cases} \frac{\gamma_p \alpha}{\gamma_p(2 + \alpha) + d + 2 \sum_{i=1}^d \beta_i} & \text{for } d_{\square} = d_{\Delta} \\ \frac{\gamma_p(\alpha + 1)}{\gamma_p(2 + \alpha) + d + 2 \sum_{i=1}^d \beta_i} & \text{for } d_{\square} = d_{f,g}. \end{cases}$$

As in the previous subsection, we obtain the same lower bound as [2] in the direct case, i.e. when $\beta_i = 0$ for all $i \in \{1, \dots, d\}$. Once again, the larger α , the easier the estimation. Moreover, smooth densities will provide a simpler classification problem.

As in Theorem 1, in the presence of noise in the variables, the rates obtained in Theorem 1 are slower. The price to pay is an additional term of the form $2 \sum_{i=1}^d \beta_i$. Nevertheless, the way where the parameters γ_p, α and the β_i interact is slightly different than for boundary fragment assumption. This is not surprising since the structure and the complexity of the problem have changed. Here γ_p denotes the regularity of $f - g$ and interacts directly with the margin parameter α .

Remark that this lower bound is valid only for $\alpha \leq 1$. Since we use in the proof of Theorem 2 an algebra based on standard Fourier analysis tools, we have to consider sufficient smooth objects. As a consequence in the lower bounds, we can check the margin assumption only for values of $\alpha \leq 1$. Nevertheless, we conjecture that this restriction is only due to technical reasons and that our result remains pertinent for all $\alpha, \gamma \in \mathbb{R}$. In particular, an interesting direction is to consider a wavelet basis which provides an isometric wavelet transforms in L^2 in order to obtain the desired lower bound in the general case.

3 Upper bounds

3.1 Estimation of G_K^*

In the free-noise case ($\epsilon_i^{(j)} = 0$ for all $i \in \{1, \dots, d\}, j \in \{1, 2\}$), we deal with two samples $(X_1^{(1)}, \dots, X_n^{(1)})$, $(X_1^{(2)}, \dots, X_m^{(2)})$ having respective densities f and g . A standard way to estimate $G_K^* = \{x \in K : f(x) \geq g(x)\}$ is to estimate $R_K(\cdot)$ thanks to the data. For all $G \subset K$, the risk $R_K(G)$ can be estimated by the empirical risk defined in (1.5). Then the Bayes classifier G_K^* is estimated by $\hat{G}_{n,m}$ defined as a minimizer of the empirical risk (1.5) over a given family of sets \mathcal{G} . We know for instance from [24] that the estimator $\hat{G}_{n,m}$ reaches the minimax rates of convergence of Theorem 1 for $\beta = 0$ when $\mathcal{G} := \mathcal{G}(\gamma, L)$ corresponds to the set of boundary fragments with $\gamma > d - 1$. For larger set $\mathcal{G}(\gamma, L)$, as proposed in [24], the minimization can be restricted to an δ -net of $\mathcal{G}(\gamma, L)$. With an additional assumption over the approximation power of this δ -net, the same minimax rates can be achieved in a subset of $\mathcal{G}(\gamma, L)$.

If we consider complexity assumptions related to the smoothness of $f - g$, we can show coarsely with [2] that an hybrid plug-in/ERM estimator reaches the minimax rates of convergence of Theorem 2 in the free-noise case. The principle of the method is to consider the empirical minimization (1.5) over a particular class \mathcal{G} based on plug-in type decision sets. More precisely, following [2] for classification, we can minimize in the direct case the empirical risk over a class \mathcal{G} of the form:

$$\mathcal{G} = \{\{f - g \geq 0\}, f - g \in \mathcal{N}_{n,m}\},$$

where $\mathcal{N}_{n,m}$ is an δ -net over the class of densities, and where $\delta := \delta_n$ is well chosen. With such a procedure, minimax rates can be obtained with no restriction over the parameter γ_p, α and d .

In noisy discriminant analysis, ERM estimator (1.6) is not available since we only observe noisy samples. The probability densities of the samples w.r.t. the Lebesgue measure are respectively convolution $(f\mu) * \eta$ and $(g\mu) * \eta$ and then classical ERM principle fails since:

$$\frac{1}{2n} \sum_{i=1}^n \mathbf{1}_{\{Z_i^{(1)} \in G^C\}} + \frac{1}{2m} \sum_{i=1}^m \mathbf{1}_{\{Z_i^{(2)} \in G\}} \xrightarrow[n \rightarrow \infty]{a.s.} \frac{1}{2} \left[\int_{K/G} (f \cdot \mu) * \eta(x) dx + \int_G (g \cdot \mu) * \eta(x) \right] \neq R_K(G).$$

Hence, we have to add a deconvolution step to the classical ERM estimator. In this context, we can construct a deconvolution kernel provided that the noise has a nonnull Fourier transform, as expressed in the *Noise Assumption*. This is rather classical in the inverse problem literature (see e.g. [14], [8], [10] or [26]). With such an assumption, we are able to construct a deconvoluting kernel as follows.

Let $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j : \mathbb{R}^d \rightarrow \mathbb{R}$ be a d -dimensional function defined as the product of d uni-dimensional function \mathcal{K}_j . The properties of \mathcal{K} leading to satisfying upper bound (depending on the considered complexity assumption) will be precised later on. Then if we denote by $\lambda = (\lambda_1, \dots, \lambda_d)$ a set of (positive) bandwidths and by $\mathcal{F}[\cdot]$ the Fourier transform, we define \mathcal{K}_η as

$$\begin{aligned} \mathcal{K}_\eta &: \mathbb{R}^d \rightarrow \mathbb{R} \\ x &\mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1} \left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)} \right]. \end{aligned} \tag{3.1}$$

In this context, for all $G \subset K$, the risk $R_K(G)$ can be estimated by

$$R_{n,m}^\lambda(G) = \frac{1}{2} \left[\frac{1}{n} \sum_{i=1}^n h_{G^C, \lambda}(Z_i^{(1)}) + \frac{1}{m} \sum_{i=1}^m h_{G, \lambda}(Z_i^{(2)}) \right],$$

where for a given $z \in \mathbb{R}$,

$$h_{G,\lambda}(z) := \int_G \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) dQ(x). \quad (3.2)$$

In the following, we propose to study ERM estimators defined as

$$\hat{G}_{n,m}^\lambda = \arg \min_{G \in \mathcal{G}} R_{n,m}^\lambda(G), \quad (3.3)$$

where the parameter $\lambda \in \mathbb{R}_+^d$ has to be chosen explicitly. It is important to note that in (3.2) $h_{G,\lambda}$ depends on Q . Hence, the measure Q needs to be known a priori. It differs from the direct case where the empirical risk is independent of the nature of Q . Here functions $h_{G,\lambda}$ in equation (3.2) are at the core of the upper bounds. In particular, remark that following the pioneering's works of Vapnik (see [33]), we have

$$\begin{aligned} R_K(\hat{G}_{n,m}^\lambda) - R_K(G^*) &\leq R_K(\hat{G}_{n,m}^\lambda) - R_{n,m}^\lambda(\hat{G}_{n,m}^\lambda) + R_{n,m}^\lambda(G_K^*) - R_K(G^*), \\ &\leq R_K^\lambda(\hat{G}_{n,m}^\lambda) - R_{n,m}^\lambda(\hat{G}_{n,m}^\lambda) + R_{n,m}^\lambda(G_K^*) - R_K^\lambda(G^*) \\ &\quad + (R_K - R_K^\lambda)(\hat{G}_{n,m}^\lambda) - (R_K - R_K^\lambda)(G_K^*) \\ &\leq \sup_{G \in \mathcal{G}} |R_K^\lambda - R_{n,m}^\lambda|(G - G_K^*) + \sup_{G \in \mathcal{G}} |R_K^\lambda - R_K|(G - G_K^*), \end{aligned} \quad (3.4)$$

where $R_K^\lambda(\cdot)$ corresponds to the expectation of $R_{n,m}^\lambda(\cdot)$. As a result, to get risk bounds, we have to deal with two opposing terms, namely a so-called variability term

$$\sup_{G \in \mathcal{G}} |R_K^\lambda - R_{n,m}^\lambda|(G - G_K^*), \quad (3.5)$$

and a bias term (since $\mathbb{E}R_{n,m}^\lambda(G) \neq R_K(G)$) of the form:

$$\sup_{G \in \mathcal{G}} |R_K^\lambda - R_K|(G - G_K^*). \quad (3.6)$$

The variability term (3.5) gives rise to the study of increments of empirical process. In this paper this control is based on entropy conditions and uniform concentration inequalities which are inspired by results presented for instance in [32] or [16]. The main novelty here is that in the noisy case, empirical processes are indexed by a class of functions which depends on the smoothing parameter λ . The bias term (3.6) is controlled by taking advantages of the properties of \mathcal{G} and of the assumptions on the kernel \mathcal{K} . Indeed, it can be related to the standard bias term in non parametric density estimation with more or less technicalities, according to the smoothness assumption (boundary fragments or plug-in type). This bias term is inherent to the proposed estimation procedure and its control is a cornerstone of the upper bounds.

The choice of λ will be a trade off between the two opposing terms (3.5) and (3.6). Small $\lambda > 0$ leads to complex functions $h_{G,\lambda}$ and blast the variance term whereas (3.6) vanishes when λ tends to zero. The kernel \mathcal{K} has to be chosen in order to take advantage of the different conditions on G_K^* . This choice will be operated according to the following definition.

Definition We say that \mathcal{K} is a kernel of order $l \in \mathbb{N}^*$ with respect to Q if and only if:

- $\int_K \mathcal{K}(u) dQ(u) = 1 \ \forall \ j = 1, \dots, d.$
- $\int_K u_j^k \mathcal{K}(u) dQ(u) = 0 \ \forall \ k = 1, \dots, l, \ \forall \ j = 1, \dots, d.$
- $\int_K |u_j|^l |\mathcal{K}(u)| dQ(u) < \infty, \ \forall \ j = 1, \dots, d.$

In addition to this definition, we will require that the deconvolution kernel is convenient for the noise η through the following assumption. Such an assumption is rather standard and is for instance satisfied if the kernel \mathcal{K} has a compactly supported Fourier transform (see the proof in the Appendix) and even under a polynomial decreasing behavior of $|\mathcal{F}[\mathcal{K}_\eta](t)|$.

Kernel Assumption. *The Kernel \mathcal{K} is such that*

$$\sup_{t \in \mathbb{R}^d} |\mathcal{F}[\mathcal{K}_\eta](t)| \leq C \prod_{i=1}^d \lambda_i^{-\beta_i},$$

for some positive constant C .

The two following subsections propose to study deconvolution ERM estimator (3.3) and give asymptotic rates of convergence for particular choices of λ . Under the margin assumption, fast and optimal rates are stated depending on the complexity assumption considered: the *Boundary fragment assumption* or the *Plug-in assumption*.

3.2 Upper bound for the plug-in assumption

We first point out that for the sake of coherence, we will not study plug-in rules in this paper, although a study similar to [2] could be managed. Since our aim is to establish minimax rates of convergence under two different complexity assumptions, we focus on the same ERM type estimators of the form (3.3).

For all $\delta > 0$, using the notion of entropy (see for instance [32]) for Hölderian function on compact sets, we can find a δ -network \mathcal{N}_δ on $\Sigma'(\gamma_p, L)$ such that

- $\log(\text{card}(\mathcal{N}_\delta)) \leq A\delta^{-d/\gamma_p}$
- For all $h_0 \in \Sigma'(\gamma_p, L)$, we can find $h \in \mathcal{N}_\delta$ such that $\|h - h_0\|_\infty \leq \delta$.

In the following, we associate to each $\nu := f - g \in \Sigma'(\gamma_p, L)$, a set $G_\nu = \{x : \nu(x) \geq 0\}$. Under the plug-in assumption, our ERM estimator will then be defined as

$$\hat{G}_{n,m} = \arg \min_{\nu \in \mathcal{N}_\delta} R_{n,m}^\lambda(G_\nu), \quad (3.7)$$

where $\delta = \delta_n$ has to be chosen carefully. This procedure has been introduced in the direct case by [2] and referred as an hybrid Plug-in/ERM procedure. The following theorem describes the performances of $\hat{G}_{n,m}$.

Theorem 3 *Let $\mathcal{F} = \mathcal{F}_{\text{plug}}$ and $\hat{G}_{n,m}$ the set introduced in (3.7) with*

$$\lambda_i = n^{-\frac{1}{\gamma_p(2+\alpha)+2\sum_{i=1}^d \beta_i + d}}, \quad \forall i \in \{1, \dots, n\}, \quad \text{and} \quad \delta_n = \left(\frac{\prod_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \right)^{\frac{2}{2/\gamma_p + 2 + \alpha}}.$$

Suppose that the noise assumption is satisfied with $\beta_i > 1/2, \forall i = 1, \dots, d$. Consider a kernel \mathcal{K}_η defined as in (3.1) where $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j$ is a kernel of order $\lfloor \gamma \rfloor$ with respect to Q , which satisfies the kernel assumption. Then

$$\lim_{n \rightarrow +\infty} \sup_{(f,g) \in \mathcal{F}_{\text{plug}}} (n \wedge m)^{\tau'_d(\alpha, \beta, \gamma)} d(\hat{G}_n, G_K^\star) < +\infty,$$

where

$$\tau'_d(\alpha, \beta, \gamma) = \begin{cases} \frac{\gamma_p \alpha}{\gamma_p(2 + \alpha) + d + 2 \sum_{i=1}^d \beta_i} & \text{for } d = d_\Delta \\ \frac{\gamma_p(\alpha + 1)}{\gamma_p(2 + \alpha) + d + 2 \sum_{i=1}^d \beta_i} & \text{for } d = d_{f,g}. \end{cases}$$

Theorem 3 validates the lower bounds of Theorem 1. Deconvolution ERM are minimax optimal over the class $\mathcal{F}_{\text{plug}}$.

These optimal rates are characterized by the tail behavior of the characteristic function of the error distribution η . We only consider the ordinary smooth case whereas straightforward modifications leads to low rates of convergence in the super smooth case.

Here fast rates are proposed provided that $\alpha\gamma > d + \sum \beta_i$. However it is important to note that large values of both α and γ corresponds to very restrictive situations. In this case the margin parameter is high whereas the behavior of $f - g$ is smooth, which seems to be contradictory (see the related discussion in [2]).

3.3 Upper bound for the boundary fragment assumption

For the sack of concision, we propose to restrict the set of all possible regularities γ_f in $\mathcal{G}(\gamma_f, L)$ to $\gamma_f > d - 1$. It allows us to control the bracketing entropy of $\mathcal{G}(\gamma_f, L)$ with a parameter $\rho = \frac{d-1}{\gamma} < 1$. Hence, the construction of the ERM estimator can be directly (at least from a theoretical point of view) performed on this set and leads to the estimator:

$$\hat{G}_n = \arg \min_{G \in \mathcal{G}(\gamma, L)} R_n^\lambda(G). \quad (3.8)$$

Nevertheless, one may also define our ERM estimator on a network in a practical purpose, without significant change in the following results.

Theorem 4 below describes the performances of \hat{G}_n for the boundary fragment assumption. It seems to highlight a difficulty to get minimax results in this setting and do not entirely validates the lower bounds of Theorem 1.

Theorem 4 *Let $\mathcal{F} = \mathcal{F}_{\text{frag}}$, $G_K^* \in \mathcal{G}(\gamma, L)$ with $\gamma > d - 1$ and $\hat{G}_{n,m}$ the set introduced in (3.8). Suppose the noise assumption is satisfied with $\beta_d \geq 1/2$. Conside a kernel \mathcal{K}_η defined as in (3.1) satysfying the Kernel assumption and such that $\mathcal{K}_{d-1} = \prod_{j=1}^{d-1} \mathcal{K}_j$ is a kernel of order $\lfloor \gamma_b \rfloor$ with respect to the Lebesgue measure. Then, for all $n \in \mathbb{N}$, we have*

$$\mathbb{E}d_{f,g}(\hat{G}_n, G_K^*) \lesssim \left(\frac{\prod_{j=1}^d \lambda_j^{-\beta_j}}{\sqrt{n}} \right)^{\frac{2(\alpha+1)\gamma}{\gamma(\alpha+2)+(d-1)\alpha}} + \sup_{G \in \mathcal{G}(\gamma, L)} |R_K^\lambda - R_K|(G).$$

In addition, if

$$\lambda_i = n^{-\frac{\alpha}{\gamma(2+\alpha)+2\alpha\gamma \sum_{i=1}^{d-1} \beta_i + 2\alpha\beta_{d+1}}}, \quad \forall i \in \{1, \dots, d-1\}, \text{ and } \lambda_d = \lambda_1^\gamma,$$

then

$$\mathbb{E}d_\Delta(\hat{G}_n, G_K^*) \lesssim n^{-\tau_d(\alpha, \beta, \gamma)} + r_n(\alpha, \lambda, G_K^*), \quad (3.9)$$

where $n^{-\tau_d(\alpha, \beta, \gamma)}$ is the optimal minimax rate of Theorem 2 and $r_n(\alpha, \lambda, G_K^*)$ is a additional term defined as:

$$r_n(\alpha, \lambda, G_K^*)$$

$$= \left(\int |f - g| |\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}| - \int (f - g)(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}) \right)^{\frac{\alpha}{\alpha+1}}.$$

Theorem 4 underlines a lack of optimality of ERM estimator \hat{G}_n for the Hölder boundary fragment assumption. It could be explain as follows.

The first assertion of Theorem 4 deals with the excess risk of the procedure. As a result, in this case using the series of inequalities (3.4), it is straightforward to get the first assertion with a modified version of Lemma 1 in [24] applied to the noisy setting. However a control of the bias term is not yet available. To deal with a boundary fragment's type assumption, we have to write the bias term using Fubini as follows:

$$R_K^\lambda(G) - R_K(G) = \int (f - g)(z) \int K(u)[\mathbf{1}_G(z) - \mathbf{1}_G(z - \lambda.u)] du dz.$$

The presence of $f - g$ in the above integral seems problematic and an assumption about the behavior of $(f - g)$ at the boundary of G_K^* seems to be necessary to reach the minimax results of Theorem 1.

To avoid the presence of $(f - g)$ in the bias term, the second assertion of Theorem 4 proposes to control $d_\Delta(\hat{G}_n, G_K^*)$. In this case, it is possible to control a bias term as follows:

$$\int \mathbf{1}_G - h_{G,\lambda} = \int K(u)[\mathbf{1}_G(z) - \mathbf{1}_G(z - \lambda.u)] du dz \leq \sum_{i=1}^{d-1} \lambda_i^\gamma + \lambda_d,$$

and to approach the minimax rates of Theorem 1 thanks to an optimal choice for λ . However in this case a residual term appears in the upper bound and the main problem is that any satisfying bound on this residual term is, up to our knowledge available. The minimax optimality of \hat{G}_n remains an open problem.

4 Conclusion

We have provided in this paper minimax rates of convergence in the framework of smooth discriminant analysis with error in variables. We consider two different assumptions over the complexity of the hypothesis space: plug-in type assumptions or boundary fragments. In the presence of plug-in type assumptions, we have proved minimax optimality reached by Deconvolution ERM. These optimal rates are fast rates (faster than $n^{-\frac{1}{2}}$) when $\alpha\gamma > d + \sum_{i=1}^d \beta_i$ and generalize the result of [2]. As shown in Table 1, the influence of the noise ϵ can be compared with standard results in regression and density estimation with errors in variables of [14, 15] using kernel deconvolution estimators.

	Density estimation	Goodness-of-fit testing	Classification
Direct case ($\epsilon = 0$)	$n^{-\frac{2\gamma}{2\gamma+1}}$	$n^{-\frac{2\gamma}{4\gamma+1}}$	$n^{-\frac{\gamma(\alpha+1)}{\gamma(\alpha+2)+d}}$
Errors in variables	$n^{-\frac{2\gamma}{2\gamma+2\beta+1}}$	$n^{-\frac{2\gamma}{4\gamma+4\beta+1}}$	$n^{-\frac{\gamma(\alpha+1)}{\gamma(\alpha+2)+2\beta+d}}$
Regularity assumptions	$f \in \Sigma(\gamma, L)$ $ \mathcal{F}[\eta](t) \sim t ^{-\beta}$	$f \in W(s, L)$ $ \mathcal{F}[\eta](t) \sim t ^{-\beta}$	$f - g \in \Sigma(\gamma, L)$ $ \mathcal{F}[\eta_i](t) \sim t ^{-\beta_i} \forall i$

Table 1. Optimal rates of convergence in pointwise L^2 -risk in density estimation (see [14]), optimal separation rates for goodness-of-fit testing on Sobolev spaces $W(s, L)$ (see e.g. [8]) and the result of the paper in smooth discriminant analysis.

In the presence of boundary fragments assumptions, we state a lower bound which generalizes the lower bound of the direct case of [24]. However Deconvolution ERM does not reach this lower bound. As a result, an open problem is to find the minimax optimal rate of convergence in the presence of noise under boundary fragments assumptions. A possible way is to find a classifier reaching the lower bound of Theorem 1. An interesting direction for this purpose

could be to consider convex loss functions in the spirit of [5]. If we take a look at Theorem 4, standard ERM with hard loss suffers from a lack of regularity. Considering for instance SVM type loss, it could be possible to control the bias term in the Deconvolution ERM using the Hölder regularity of the boundary. The robustness of SVM with respect to noise in variables could be an interesting future work. However, this paper seems to highlight that at the first glance plug-in type assumptions are more adapted to the presence of noise in classification.

We conclude this discussion by some words on adaptation. It is important to note that considering the estimation procedure of this paper, we are faced to two different problems of model selection or adaptation. First of all the bandwidths proposed in this paper clearly depend on parameters which may be unknown a priori (e.g. the margin α or the regularity of the boundary γ). In this sense, adaptation algorithms should be investigated to choose automatically λ to balance the bias term and the variance term. The second step of adaptation would be to consider a family of nested $(\mathcal{G}_k) \subset \mathcal{G}$ and to choose the model which balance the approximation term and the estimation term. This could be done using for instance penalization techniques, such as [31] or [19].

This work can be considered as a first attempt into the study of risk bounds in classification with errors in variables. It can be extended in many directions. Naturally the first extension will be to state the same kind of result in classification. Another natural direction would be to consider more general complexity assumptions for the hypothesis space \mathcal{G} . In the free noise case, [4] proposes to deal with Local Rademacher complexities. It allows to consider many hypothesis spaces, such as VC class of sets, kernel classes (see [27]) or even Besov spaces (see [23]). Another advantage of considering Rademacher complexities is to develop data-dependent complexities to deal with the problem of model selection (see [19, 3]) and to deal with the problem of non-unique solution of the empirical minimization.

Into the direction of statistical inverse problem, there are also many directions of study. A natural direction for applications would be to consider unknown density η for the random noise ϵ . this is a well known issue in the inverse problem litterature to deal with unknown operator of inversion. Another natural extension will be to consider general linear compact operator $A : f \mapsto Af$ to generalize the case of deconvolution. In this case, ERM estimators based on standard regularization methods from the inverse problem litterature (see [13]) appear as good candidates. This could be the material of future works.

In this paper, a classifier $G : \mathcal{X} \rightarrow \mathcal{Y}$ is always identified with a subset of \mathbb{R}^d . Our aim is then to estimate the set G_K^* from the noisy observations (1.3). In particular, the main goal is not only to provide a good classifier but also to understand the relationship between the spatial position of an input $X \in \mathbb{R}^d$ and its affiliation to one of the candidate densities. One could alternatively try to provide the best classifier for a noisy input Z from a noisy training set. These two problems are certainly comparable, although a rigorous comparison of the two framework and the respective error of classification should be done. We mention for instance [21] or [22] for a related discussion in a goodness-of-fit purpose. This could be the core of a future work, but it requires the preliminary study provided in this paper.

5 Proofs

In this section, with a slight abuse of notations, $C, c, c' > 0$ denotes generic constants that may vary from line to line, and even in the same line. The notation $a \approx b$ (resp. $a \lesssim b$) means that there exists generic constants $C, c > 0$ such that $ca \leq b \leq Ca$ (resp. $a \leq Cb$).

5.1 Proof of Theorem 1

The proof starts as in [24] but then uses some arguments which are specific to the inverse problem literature (see for instance [8] or [26]).

Let \mathcal{F}_1 a finite class of densities and g_0 a fixed density such that $(f, g_0) \in \mathcal{F}_{\text{frag}}$ for all $f \in \mathcal{F}_1$. The composition of \mathcal{F}_1 and the value of g_0 will be precised later on. Then, for all estimator $\hat{G}_{n,m}$ of the set G_K^* , we have

$$\begin{aligned} \sup_{(f,g) \in \mathcal{F}_{\text{frag}}} \mathbb{E}_{f,g} d_{\Delta}(\hat{G}_{n,m}, G_K^*) &\geq \sup_{(f,g_0), f \in \mathcal{F}_1} \mathbb{E}_{f,g} d_{\Delta}(\hat{G}_{n,m}, G_K^*), \\ &\geq \mathbb{E}_{g_0} \left[\frac{1}{\#\mathcal{F}_1} \sum_{f \in \mathcal{F}_1} \mathbb{E}_f \left\{ d_{\Delta}(\hat{G}_{n,m}, G_K^*) | X_1^{(2)}, \dots, X_m^{(2)} \right\} \right]. \end{aligned} \quad (5.1)$$

5.1.1 Construction of \mathcal{F}_1

Concerning the density g_0 , we deal with the uniform density on $[0, 1]^2$, i.e.

$$g_0(x) = \mathbf{1}_{\{x \in [0,1]^2\}}, \forall x \in \mathbb{R}^2.$$

Now, we have to define the class \mathcal{F}_1 . First, we consider a function φ infinitely differentiable defined on \mathbb{R} such that $\text{supp}(\varphi) = [-1, 1]$, $\varphi(t) \geq 0$ for all $t \in \mathbb{R}$ and $\|\varphi\|_{\infty} = \varphi(0) = 1$. Let $M \geq 2$ an integer which will be allowed to depend on n and $\tau > 0$ a positive constant. Then, for all $j \in \{1, \dots, M\}$, we set

$$\varphi_j(t) = \tau M^{-\gamma} \varphi \left(M \left[t - \frac{2j-1}{M} \right] \right), \quad \forall t \in \mathbb{R}.$$

For all $\omega \in \{0, 1\}^M$ and all $t \in \mathbb{R}$, we define

$$b(t, \omega) = \frac{1}{2} + \sum_{j=1}^M \omega_j \varphi_j(t).$$

In the specific case where $\omega_j = 1$ for all $j \in \{1, \dots, M\}$, we write $b(t, \mathbf{1})$. Then, let b_0 and C^* positive constants which will be precised later on. We define the function $f_0 : \mathbb{R}^2 \rightarrow \mathbb{R}$ as $f_0(x) = 0$ for all $x \notin [0, 1]^2$ and

$$f_0(x) = \begin{cases} 1 + 2\eta_0, \forall x_2 \in [0, 1/2], \\ 1 - \eta_0 - b_0, \forall x_2 \in [b(x_1, \mathbf{1}), 1], \\ 1 + \left(\frac{b(x_1, \mathbf{1}) - x_2}{c_2} \right)^{1/\alpha} - C^* M^{-\gamma/\alpha}, \forall x_2 \in [1/2, b(x_1, \mathbf{1})], \end{cases}$$

where $C^* = 3/2 \cdot (\tau/c_2)^{1/\alpha}$ and $b_0 > 0$ is such that $\int f_0(x) dx = 3/4$. The condition on C^* ensures that $f_0(x) < 1$ for all $x_2 \in [1/2, b(x_1, \mathbf{1})]$. We will also use the function f_1 defined as

$$f_1(x) = \begin{cases} 0, \forall x \in [0, 1]^2, \\ \frac{\mathcal{C}_1}{(1+x_2)^2 \cdot (1+x_1)^2}, \forall x \notin [0, 1]^2, \end{cases}$$

where \mathcal{C}_1 is such that $\int f_1(x) dx = 1/4$. Finally, the set \mathcal{F}_1 will be defined as

$$\mathcal{F}_1 = \{f_{\omega}, \omega \in [0, 1]^M\},$$

where for a given $\omega \in \{0, 1\}^M$,

$$f_\omega(x) = f_0(x) + f_1(x) + \sum_{j=1}^M \omega_j \rho_j(x). \quad (5.2)$$

for some functions $(\rho_j)_{j=1\dots M}$ which are explicated below. In order to complete the construction of the set \mathcal{F}_1 , we have to provide a precise definition of the ρ_j and to prove that the f_ω define probability density functions for all $\omega \in \{0, 1\}^M$.

We first start with the construction of the ρ_j . For all $x \in \mathbb{R}$, let $\rho : \mathbb{R} \rightarrow [0, 1]$ the function defined as

$$\rho(x) = \frac{1 - \cos(x)}{\pi x^2}, \quad \forall x \in \mathbb{R},$$

with associate Fourier transform $\mathcal{F}[\rho](t) = (1 - |t|)_+$. In particular, $\text{supp } \mathcal{F}[\rho] = [-1, 1]$. For all $j \in \{1, \dots, M\}$ and $x_2 \in \mathbb{R}$, introduce

$$\rho_{(2)}(x_2) = \cos\left(\frac{x_2 - 1/2(1 + \tau M^{-\gamma})}{3/2\pi^{-1}\tau M^{-\gamma}}\right) \rho\left(\frac{x_2 - 1/2(1 + \tau M^{-\gamma})}{3\pi^{-1}\tau M^{-\gamma}}\right). \quad (5.3)$$

By the same way, for all $j \in \{1, \dots, M\}$, we define

$$\rho_{j,(1)}(x_1) = \cos\left[\frac{\pi}{3}\left(\frac{x_1 - j/M}{M^{-1}}\right)\right] \rho\left[\frac{\pi}{6}\left(\frac{x_1 - j/M}{M^{-1}}\right)\right]. \quad (5.4)$$

Then, for all $j \in \{1, \dots, M\}$ and $x = (x_1, x_2) \in [0, 1]^2$, we set

$$\rho_j(x) = c^*(\tau M^{-\gamma})^{1/\alpha} \rho_{(2)}(x_2) \rho_{j,(1)}(x_1), \quad (5.5)$$

for some constant c^* explicated below.

Now, we prove that the f_ω introduced in (5.2) define density functions. First, remark that

$$\sum_{j=1}^M |\rho_j(x)| \leq \begin{cases} CM^{-\gamma/\alpha} (1 + x_1)^{-2} (1 + x_2)^{-2}, & \forall x \notin [0, 1]^2, \\ CM^{-\gamma/\alpha}, & \forall x \in [0, 1]^2, \end{cases}$$

This ensures that $f_\omega \geq 0$ for all $\omega \in \{0, 1\}^M$, at least for M large enough. Then recall that both f_0 and f_1 are designed in order to guarantee that $\int (f_0 + f_1)(x) dx = 1$. Hence, we only have to show that $\int \rho_j(x) dx = 0$ for all $j \in \{1, \dots, M\}$. In fact, it is only necessary to prove that $\int \rho_{(2)}(x_2) dx_2 = 0$. First remark that $\int \rho_{(2)}(x_2) dx_2 = \int \tilde{\rho}_{(2)}(x_2) dx_2$ where $\tilde{\rho}_{(2)}(x_2) = \rho_{(2)}(x_2 + 1/2(1 + \tau M^{-\gamma}))$ for all $x_2 \in \mathbb{R}$. Then, using simple algebra

$$\mathcal{F}[\rho_{(2)}](0) = \frac{1}{2} \mathcal{F}[\rho]\left(\frac{1}{3\pi^{-1}\tau M^{-\gamma}}\right) \left(\pm \frac{1}{3/2\pi^{-1}\tau M^{-\gamma}}\right) = \frac{3}{2} \pi^{-1} \tau M^{-\gamma} \mathcal{F}[\rho](\pm 2) = 0,$$

since the support of the Fourier transform of ρ is $[-1, 1]$. Hence, for all $\omega \in \{0, 1\}^M$, f_ω is a density function.

In order to conclude the proof, we have to show that

$$(f_\omega, g_0) \in \mathcal{F}_{\text{frag}} \quad \forall \omega \in \{0, 1\}^M, \quad (5.6)$$

which allows to use the bound (5.1),

$$Q\{x \in K : |f_\omega(x) - g_0(x)| \leq \eta\} \leq c_2 \eta^\alpha \quad \forall \omega \in \{0, 1\}^M \text{ and } \forall \eta \leq \eta_0, \quad (5.7)$$

which means that the *Margin assumption* is satisfied for our test functions and that

$$\mathbb{E}_{g_0} \mathbb{E}_{f_\omega} \left\{ d_\Delta(\hat{G}_{n,m}, G_K^*) | X_1^{(2)}, \dots, X_m^{(2)} \right\} \geq C n^{-\frac{\gamma}{\gamma(\frac{2}{\alpha}+1)+2\beta_1+2\beta_2\gamma+1}}, \quad (5.8)$$

for some positive constant C .

5.1.2 Main assumptions check

We first start with the proof of (5.6). First remark that for all $j \in \{1, \dots, M\}$, the function $\rho_j(\cdot)$ is bounded from above by $CM^{-\gamma/\alpha}$ for some $C > 0$. Then, using simple algebra

$$\begin{aligned} x_2 \in [1/2; b(x_1, \mathbf{1})] &\Rightarrow \frac{1}{2} \leq x_2 \leq \frac{1}{2} + \tau M^{-\gamma}, \\ &\Rightarrow -\frac{\tau M^{-\gamma}}{2} \leq x_2 - \frac{1}{2} - \frac{\tau M^{-\gamma}}{2} \leq \frac{\tau M^{-\gamma}}{2}, \\ &\Rightarrow -\frac{\pi}{6} \leq \frac{x_2 - 1/2(1 + \tau M^{-\gamma})}{3\pi^{-1}\tau M^{-\gamma}} \leq \frac{\pi}{6}, \\ &\Rightarrow \rho_{(2)}(x_2) \geq \frac{9}{4\pi^3}. \end{aligned}$$

The same kind on minoration holds for the function $\rho_{j,(1)}$. Hence the ρ_j are uniformly bounded from below on $[1/2; b(x_1, \mathbf{1})]$. For all $\omega \in \{0, 1\}^M$ and for all $x \in [0, 1]^2$, we have

$$f_\omega(x) \geq 1 + \left(\frac{b(x, \mathbf{1}) - x_2}{c_2} \right)^{1/\alpha} \geq g_0(x), \quad \forall x_2 \in [1/2, b(x_1, \omega)],$$

for c^* large enough. This ensures that

$$\{x \in [0, 1]^2 : f_\omega(x) \geq g_0(x)\} = \{x \in [0, 1]^2 : 0 \leq x_2 \leq b(x_1, \omega)\}.$$

In order to conclude the proof of (5.6), we only have to remark that the function $b(\cdot, \omega)$ belongs to $\Sigma(\gamma, L)$ for all $\omega \in \{0, 1\}^M$, at least for M small enough.

Now, we consider the margin assumption (5.7). First, we consider the case where $\eta < [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha} < \eta_0$. Clearly, following our choices of b_0 and C^* , we have that

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow x_2 \in [1/2; b(x_1, \omega)] \Rightarrow x_2 \leq b(x_1, \omega).$$

Moreover, for all $x \in [0, 1]^2$ such that $x_2 \leq b(x_1, \omega)$, we have

$$(f_\omega - g_0)(x) = \left(\frac{b(x, \mathbf{1}) - x_2}{c_2} \right)^{1/\alpha} + \sum_{j=1}^M \omega_j \rho_j(x) - C^* M^{-\gamma/\alpha},$$

where

$$\sum_{j=1}^M \omega_j \rho_j(x) - C^* M^{-\gamma/\alpha} > 0, \quad \forall x_2 \in \left[\frac{1}{2}, b(x_1, \omega) \right].$$

Thus

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow \left(\frac{b(x, \omega) - x_2}{c_2} \right)^{1/\alpha} \leq \eta \Rightarrow x_2 \geq b(x_1, \omega) - c_2 \eta^\alpha,$$

which proves the margin assumption when $\eta < [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha}$. Now, in the case where $\eta_0 > \eta > [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha}$, we have

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow 1/2 < x_2 < b(x_1, \mathbf{1}),$$

which entails

$$Q \{x \in K : |f_\omega(x) - g_0(x)| \leq \eta\} \leq \tau M^{-\gamma} \leq c_2 \eta^\alpha.$$

This concludes this part.

5.1.3 Final minoration

Now, we can deal with the lower bound (5.8). The proof is based on classical tools which can be found for instance in [29], [24], [8] or [26]. First remark that the shape of G_K^* depends on the value of ω . For the sake of convenience, we omit the dependency with respect to this quantity. For all $\omega \in \{0, 1\}^M$, recall that

$$G_K^* = \{x \in [0, 1]^2 : f_\omega(x) \geq g_0(x)\} = \{x \in [0, 1]^2 : 0 \leq x_2 \leq b(x_1, \omega)\}.$$

Using Assouad Lemma and classical tools designed for instance in [29], we get

$$\mathbb{E} \left[d_\Delta(\hat{G}_{n,m}, G_K^*) | Y_1, \dots, Y_m \right] \geq \frac{M}{2} \|\varphi_1\|_1 \int \min[dP_{11}, dP_{10}], \quad (5.9)$$

where P_{11} denotes the law of $(Z_i^{(1)})_{i=1 \dots n}$ when the density of the $X_i^{(1)}$ is $f_{\omega_{11}}$. In the following, we will choose M in order to guarantee that the term $\int \min[dP_{11}, dP_{10}]$ is bounded from below. Consequently, the lower bound will be determined by the corresponding value of $M\|\varphi_1\|_1$. Since the observations are independent

$$\int \min[dP_{11}, dP_{10}] \geq 1 - \sqrt{(1 + \chi^2(P_1, P_0))^n - 1},$$

where $\chi^2(P_a, P_b)$ denotes the chi-square divergence between two given probability measures P_a and P_b , and P_0, P_1 are the law of the variable $Z_1^{(1)} = X_1^{(1)} + \epsilon_1^{(1)}$ when the density of the X_i is respectively $f_{\omega_{11}}$ or $f_{\omega_{10}}$. In the following, our aim is to find a satisfying upper bound for $\chi^2(P_1, P_0)$.

First, remark that we can find $\tilde{c} > 0$ such that for all $x \notin [0, 1]^2$ and all $\omega \in \{0, 1\}^M$, $f_\omega(x) \geq \tilde{c}f_1(x)$. Hence, using simple algebra, we get that

$$f_\omega * \eta(x) \geq \frac{C}{(1 + x_1^2)(1 + x_2^2)}, \quad \forall x \in \mathbb{R}^2, \quad (5.10)$$

for some $C > 0$. In the following, given f, η_1 and η_2 , we denote by $f * \eta$ the convolution product in dimension two, i.e.

$$f * \eta(x) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x_1 - y_1, x_2 - y_2) \eta_1(y_1) \eta_2(y_2) dy_1 dy_2, \quad \forall x \in \mathbb{R}^2.$$

Then, using (5.2) and (5.10),

$$\begin{aligned} \chi^2(P_1, P_0) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{\{(f_{\omega_{11}} - f_{\omega_{10}}) * \eta(x)\}^2}{f_{\omega_{11}} * \eta(x)} dx, \\ &\leq C \int_{\mathbb{R}} \int_{\mathbb{R}} (1 + x_1^2)(1 + x_2^2) \{\rho_1 * \eta(x)\}^2 dx. \end{aligned}$$

Hence

$$\begin{aligned} \chi^2(P_1, P_0) &\leq C \int_{\mathbb{R}} \int_{\mathbb{R}} \{\rho_1 * \eta(x)\}^2 dx + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_2^2 \{\rho_1 * \eta(x)\}^2 dx \\ &\quad + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_1^2 \{\rho_1 * \eta(x)\}^2 dx + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_1^2 x_2^2 \{\rho_1 * \eta(x)\}^2 dx, \\ &:= A_1 + A_2 + A_3 + A_4, \end{aligned}$$

where the ρ_j are defined in (5.5). In the following, we only consider the bound of A_1 , the other terms being controlled in the same way. We get

$$\begin{aligned}
A_1 &= C \int_{\mathbb{R}} \int_{\mathbb{R}} \{\rho_1 * \eta(x)\}^2 dx, \\
&= CM^{-2\gamma/\alpha} \int_{\mathbb{R}} \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}} \rho_{(2)}(x_2 - y_2) \rho_{j,(1)}(x_1 - y_1) \eta_1(y_1) \eta_2(y_2) dy_1 dy_2 \right\}^2 dx, \\
&= CM^{-2\gamma/\alpha} \int_{\mathbb{R}} \int_{\mathbb{R}} |\mathcal{F}[\rho_{(2)}](t_2)|^2 |\mathcal{F}[\rho_{1,(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 |\mathcal{F}[\eta_2](t_2)|^2 dt_1 dt_2, \\
&= CM^{-2\gamma/\alpha} A_{1,1} A_{1,2},
\end{aligned}$$

where

$$A_{1,1} = \int_{\mathbb{R}} |\mathcal{F}[\rho_{(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \quad A_{1,2} = \int_{\mathbb{R}} \int_{\mathbb{R}} |\mathcal{F}[\rho_{1,(2)}](t_2)|^2 |\mathcal{F}[\eta_2](t_2)|^2 dt_2,$$

and $\rho_{(1)}$, $\rho_{1,(2)}$ are respectively defined in (5.3), (5.4). We first deal with the term $A_{1,2}$. Using simple algebra, we get

$$\begin{aligned}
A_{1,2} &= \int_{\mathbb{R}} |\mathcal{F}[\rho_1^{(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \\
&= \int_{\mathbb{R}} \left| \mathcal{F} \left[\rho \left(\frac{\cdot}{3\pi^{-1}\tau M^{-\gamma}} \right) \right] \left(t_1 \pm \frac{1}{3/2\pi^{-1}\tau M^{-\gamma}} \right) \right|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \\
&= (3\pi^{-1})^2 \tau^2 M^{-2\gamma} \int_{\mathbb{R}} \left| \mathcal{F}[\rho] \left(3\pi^{-1}\tau M^{-\gamma} t_1 \pm \frac{3}{2} \right) \right|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1.
\end{aligned}$$

Then, setting $s_1 = 3\pi^{-1}\tau M^{-\gamma} t_1$ and using the *Noise assumption*, we obtain

$$\begin{aligned}
A_{1,2} &= 3\pi^{-1}\tau M^{-\gamma} \int_{\mathbb{R}} |\mathcal{F}[\rho](s_1 \pm 2)|^2 \left| \mathcal{F}[\eta_1] \left(\frac{s_1}{3\pi^{-1}\tau M^{-\gamma}} \right) \right|^2 ds_1, \\
&= 3\pi^{-1}\tau M^{-\gamma} \int_1^3 |\mathcal{F}[\rho](s_1 \pm 2)|^2 \left| \mathcal{F}[\eta_1] \left(\frac{s_1}{3\pi^{-1}\tau M^{-\gamma}} \right) \right|^2 ds_1, \\
&\leq CM^{-\gamma-2\beta_2\gamma} \int_1^3 |\mathcal{F}[\rho](s_1 \pm 2)|^2 |s_1|^{-2\beta_1} ds_1, \\
&\leq CM^{-\gamma-2\beta_2\gamma}.
\end{aligned}$$

Using a similar algebra for the term $A_{1,1}$, we obtain

$$A_{1,2} \leq CM^{-1-2\beta_1}.$$

Similar bounds are available for A_2, A_3 and A_4 since $\mathcal{F}[\rho]$ and its weak derivative are bounded by 1 and supported on $[-1; 1]$. In particular, we use the fact that for all $t \in \mathbb{R}$

$$\mathcal{F}[\rho_{1,(2)}](t) = 3\pi^{-1}\tau M^{-\gamma} \mathcal{F}[\rho](3\pi^{-1}\tau M^{-\gamma} t \pm 2),$$

and

$$\frac{d}{dt} \mathcal{F}[\rho_{1,(2)}](t) = -i(3\pi^{-1}\tau M^{-\gamma})^2 t \mathcal{F}[\rho](3\pi^{-1}\tau M^{-\gamma} t \pm 2),$$

for all t in a subset of \mathbb{R} having a Lebesgue measure equal to 1.

The above equations lead to the following upper bound:

$$\chi^2(P_1, P_0) \leq CM^{-\gamma(2/\alpha+1)-2\beta_1\gamma-2\beta_2-1}.$$

Then, $\chi^2(P_1, P_0) \leq C/n$ for some constant $C > 0$ as soon as

$$M = M_n \sim n^{\frac{1}{\gamma(2/\alpha+1)+2\beta_1+2\beta_2\gamma+1}}.$$

Finally, going back to equation (5.9), we obtain

$$\begin{aligned} \mathbb{E} \left[d_\Delta(\hat{G}_{n,m}, G_K^*) | Y_1, \dots, Y_m \right] &\geq \frac{M_n}{2} \|\varphi_1\|_1 \int \min[dP_{11}, dP_{10}], \\ &\geq CM_n \|\varphi_1\|_1, \\ &= C\tau M_n^{-\gamma} \int_0^1 \varphi_1(t) dt, \\ &\sim M_n^{-\gamma} = n^{-\frac{\gamma}{\gamma(2/\alpha+1)+2\beta_1+2\beta_2\gamma+1}}, \end{aligned}$$

which concludes the proof.

5.2 Proof of Theorem 2

The proof mixes standard lower bounds arguments coming from classification (see [1] and [2]) but then uses some techniques which are specific to the inverse problem literature (see for instance [8] or [26]).

Consider $\mathcal{F}_2 = \{f_{\vec{\sigma}}, \vec{\sigma} = (\sigma_1, \dots, \sigma_k) \in \{0, +1\}^k\}$ a finite class of densities with respect to a specific measure Q_0 and g_0 a fixed density (with respect to the same Q_0) such that $(f_{\vec{\sigma}}, g_0) \in \mathcal{F}_{\text{plug}}$ for all $\vec{\sigma} \in \{-1, +1\}^k$. The construction of $f_{\vec{\sigma}}$ as a function of $\vec{\sigma}$, the value of g_0 and the definition of Q_0 will be precised in Section 5.2.1. Then, for all estimator $\hat{G}_{n,m}$ of the set G_K^* , we have:

$$\sup_{(f,g) \in \mathcal{F}_{\text{plug}}} \mathbb{E}_{f,g} d_\Delta(\hat{G}_{n,m}, G_K^*) \geq \sup_{f \in \mathcal{F}_2} \mathbb{E}_{g_0} \left[\mathbb{E}_f \left\{ d_\Delta(\hat{G}_{n,m}, G_K^*) | Z_1^{(2)}, \dots, Z_m^{(2)} \right\} \right]. \quad (5.11)$$

In a first time, we propose a triplet $(\mathcal{F}_2, g_0, Q_0)$. Then we prove that each associated element satisfies our hypotheses. We finish the proof with a convenient lower bound for (5.11).

5.2.1 Construction of the triplet $(\mathcal{F}_2, g_0, Q_0)$

We only consider the case $d = 2$ for simplicity, whereas straightforward modifications lead to the general d -dimensional case. For g_0 , we take the constant 1 over \mathbb{R}^d :

$$g_0(x) = 1, \forall x \in \mathbb{R}^d.$$

For any $z \in \mathbb{R}^d$ and positive δ , we write in the sequel $B(z, \delta) := \{x = (x_1, \dots, x_d) : |x_i - z_i| \leq \delta\}$.

For an integer $q \geq 1$, introduce the regular grid on \mathbb{R}^d defined as:

$$G_q = \left\{ \left(\frac{2p_1 + 1}{2q}, \dots, \frac{2p_d + 1}{2q} \right), p_i \in \{0, \dots, q-1\}, i = 1, \dots, d \right\}.$$

Let $n_q(x) \in G_q$ the closest point to $x \in \mathbb{R}^d$ among points in G_q (by convention, we choose the closest point to 0 when it is non unique). Consider the partition $(\chi'_j)_{j=1, \dots, q^d}$ of $[0, 1]^d$ defined as follows: x and y belongs to the same subset if and only if $n_q(x) = n_q(y)$. Fix an integer $k \leq q^d$. For any $i \in \{1, \dots, k\}$, we define $\chi_i = \chi'_i$ and $\chi_0 = \mathbb{R}^d \setminus \cup_{i=1}^k \chi_i$ to get $(\chi_i)_{i=1, \dots, k}$ a partition of \mathbb{R}^d .

Then, we consider the measure Q_0 defined as $dQ_0(x) = \mu(x)dx$ where $\mu(x) = \mu_0(x) + \mu_1(x)$ for all $x \in \mathbb{R}^2$ with

$$\mu_0(x) = k\omega\rho(x_1 - 1/2)\rho(x_2 - 1/2) \text{ and } \mu_1(x) = (1 - k\omega)\rho(x_1 - a)\rho(x_2 - b)$$

where k, ω, a, b are constants which will be precised later on and where for all $x \in \mathbb{R}$, $\rho : \mathbb{R} \rightarrow [0, 1]$ is the function defined in the previous lower bound as

$$\rho(x) = \frac{1 - \cos(x)}{\pi x^2}, \quad \forall x \in \mathbb{R}.$$

It seems clear that the function g define a probability density w.r.t. to the measure Q_0 since $\int_{\mathbb{R}^2} \mu(x) dx = 1$.

Now, we have to define the class $\mathcal{F}_2 = \{f_{\vec{\sigma}}, \vec{\sigma}\}$. Denote by $(z_j)_{j=1, \dots, k}$ the centers of the χ'_j s. We first introduce φ as a C^∞ probability density function w.r.t. the measure Q_0 and such that

$$\varphi(x) = 1 - c^* q^{-\gamma} \quad \forall x \in [0, 1]^2.$$

Now introduce a class of function $\psi_j : \mathbb{R}^2 \rightarrow \mathbb{R}$, for $j = 1, \dots, k$ defined for any $x \in \mathbb{R}^2$ as follows:

$$\psi_j(x) = q^{-\gamma} c_\psi \rho(2\pi q(x_1 - z_1^j)) \rho(2\pi q(x_2 - z_2^j)) \cos(4\pi q(x_1 - z_1^j)) \cos(4\pi q(x_2 - z_2^j)),$$

The class $(\psi_j)_j$ is specific to the noisy case and the inverse problem literature (see [8] and [26]), and mimics the construction provided in Theorem 1. Recall that ρ satisfies $\mathcal{F}[\rho](t) = (1 - |t|)_+$, and will allow us to take advantages of the regularity assumption over η in the *noise assumption*.

With such notations, for any $\vec{\sigma} \in \{0, 1\}^d$, we define:

$$f_{\vec{\sigma}}(x) = \varphi(x) + \sum_{l=1}^k \sigma_l \psi_l(x), \quad \forall x \in \mathbb{R}^2.$$

Now we have to check that this choice of \mathcal{F}_2, g_0 and Q_0 provides the margin assumption and that the complexity assumption hold true.

5.2.2 Properties of the triplet $(\mathcal{F}_2, g_0, Q_0)$

In a first time, we prove that the $f_{\vec{\sigma}}$ define probability density function w.r.t. the measure Q_0 . Let $\vec{\sigma} \in \{0, 1\}^k$. Remark that, considering the case $d = 1$ w.l.o.g:

$$\int_{\mathbb{R}} \psi_l(x) \mu_0(x) dx = \mathcal{F}[\psi_l \mu_0](0) = c_\psi q^{-\gamma} \mathcal{F}[\rho(2\pi q \cdot) \mu_0(\cdot)](\pm 4\pi q) = c_\psi q^{-\gamma} k \omega \mathcal{F}[\rho] * \mathcal{F}[\rho(2\pi q \cdot)](\pm 4\pi q).$$

Then, since

$$\mathcal{F}[\rho(2\pi q \cdot)](t) = \frac{1}{2\pi q} \mathcal{F}[\rho]\left(\frac{t}{2\pi q}\right) \quad \forall t \in \mathbb{R},$$

and

$$\mathcal{F}[\rho(2\pi q \cdot)](t) \neq 0 \Leftrightarrow -1 < \frac{t}{2\pi q} < 1 \Leftrightarrow -2\pi q < t < 2\pi q,$$

we get

$$\text{supp} \mathcal{F}[\rho] * \mathcal{F}[\rho(2\pi q \cdot)] = [-2\pi q - 1; 2\pi q + 1] \text{ and } \int_{\mathbb{R}} \psi_l(x) \mu_0(x) dx = 0. \quad (5.12)$$

This proves the desired result.

Concerning the regularity, $f_{\vec{\sigma}} \in \Sigma(\gamma, L)$ for q large enough since $f_{\vec{\sigma}}$ can be written as $q^{-\gamma} F_0(x)$ where F_0 is infinitely differentiable.

In order to conclude this part, we only have to prove that the margin hypothesis is satisfied for all the couples $(f_{\vec{\sigma}}, g)$. Concerning the parameters k and ω we will use the following asymptotics

$$\begin{cases} k\omega = O(q^{-\alpha\gamma}), \\ k = q^d, \\ w = q^{-\alpha\gamma-d}. \end{cases}$$

Then, we will distinguish two different cases concerning the possible value of t . The first case concerns the situation where $C_1 q^{-\gamma} < t < t_0$ for some constant C_1 . Then, thanks to the construction of μ_0 :

$$\mu_0 \left\{ x \in [0, 1]^d : |f_{\vec{\sigma}}(x) - g(x)| < t \right\} \leq \int_{[0, 1]^2} \mu_0(x) dx \leq k\omega \leq Cq^{-\alpha\gamma} \leq Ct^\alpha.$$

Now, we consider the case where $t < C_1 q^{-\gamma}$. We have in dimension $d = 2$ for simplicity, $\forall \sigma \in \{0, 1\}^k$:

$$\begin{aligned} \mu_0 \left\{ x \in [0, 1]^2 : |(f_\sigma - g)(x)| \leq t \right\} &= \int_{[0, 1]^2} k\omega \mathbf{1}_{|(f_\sigma - g)(x)| \leq t} dx \leq k\omega \sum_{j=1}^k \int_{\chi_j} \mathbf{1}_{|(f_\sigma - g)(x)| \leq t} dx \\ &\leq k^2 \omega \text{Leb} \{ x \in \chi_1 : |(f_\sigma - g)(x)| \leq t \}, \end{aligned} \quad (5.13)$$

where without loss of generality, we suppose that $\sigma_1 = 1$.

Last step is to control the Lebesgue measure of the set $W_1 = \{x \in \chi_1 : |(f_\sigma - g)(x)| \leq t\}$. Since $f_\sigma - g = \sum_{j=1}^k \sigma_j \psi_j - C^* q^{-\gamma}$, we have:

$$W_1 \subset \{x \in \chi_1 : \left| \sum_{j=1}^k \sigma_j \psi_j(x) \right| \leq t\} \subset \{x \in \chi_1 : |\psi_1(x)| \leq t\} := W'_1,$$

noting that $\forall j' \neq j$, $\text{sign} \psi_j = \text{sign} \psi_{j'}$. We hence have to control the size of W'_1 . The idea is to approximate ψ_x at each $x \in W'_1$ by a Taylor polynomial of order 1 at $z^x := \arg \min_{z: \psi_1(z)=0} \|x - z\|$ as follows:

$$\psi_1(x) = \psi_1(z_x) + \nabla \psi_1(z_1^x, z_2^x) \cdot (x - z^x) + o(\|x - z\|).$$

Hence we have by construction, since $\forall x \in W'_1$, there exists $i \in \{1, 2\} : x_i = z_i^x$:

$$\begin{aligned} \text{Leb}(W'_1) &\leq \text{Leb} \{ x \in \chi_1 : |\psi_1(z_x) + \nabla \psi_1(z_1^x, z_2^x) \cdot (x - z^x)| \leq t \} \\ &\leq c \text{Leb} \{ x \in \chi_1 : q q^{-\gamma} |x_1 - z_1^x| \leq t \} \\ &\leq c \frac{t}{q^2 q^{-\gamma}}. \end{aligned}$$

Gathering with (5.13), we hence get, for $t < C_1 q^{-\gamma}$, provided that $\alpha \geq 1$:

$$\begin{aligned} \mu_0 \left\{ x \in [0, 1]^2 : |(f_\sigma - g)(x)| \leq t \right\} &\leq c k^2 \omega \frac{t}{q^2 q^{-\gamma}} \\ &\leq c k \omega \frac{t}{q^{-\gamma}} = c q^{\gamma(1-\alpha)} t^\alpha t^{1-\alpha} \leq c' t^\alpha. \end{aligned}$$

5.2.3 Final minoration

Suppose without loss of generality that $n \leq m$. Now we argue as in [1] (Assouad Lemma for classification) and introduce ν , the distribution of a Bernoulli variable ($\nu(\sigma = 1) = \nu(\sigma = 0) =$

1/2). Then, denoting by $\mathbb{P}_{\vec{\sigma}}^{\otimes n}$ the law of $(Z_1^{(1)}, \dots, Z_n^{(1)})$ when $f = f_{\vec{\sigma}}$, we get

$$\begin{aligned}
& \sup_{\vec{\sigma} \in \{-1, +1\}} \mathbb{E}_f \left\{ d_{\Delta}(\hat{G}_{n,m}, G_K^*) | Z_1^{(2)}, \dots, Z_m^{(2)} \right\} \\
& \geq \mathbb{E}_{\nu^{\otimes k}} \mathbb{E}_{f_{\vec{\sigma}}} d_{\Delta}(\hat{G}_{n,m}, G_K^*), \\
& \geq \mathbb{E}_{\nu^{\otimes k}} \mathbb{E}_{f_{\vec{\sigma}}} \sum_{j=1}^k \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m} \Delta G_K^*) Q_0(dx), \\
& = \mathbb{E}_{\nu^{\otimes(k-1)}} \sum_{j=1}^k \int_{\Omega} \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^*) Q(dx) \mathbb{P}_{\vec{\sigma}}^{\otimes n}(d\omega) \\
& \geq \mathbb{E}_{\nu^{\otimes(k-1)}} \sum_{j=1}^k \int_{\Omega} \left[\frac{\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}}{\mathbb{P}_{\vec{\sigma}_j}^{\otimes n}} \wedge \frac{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}}{\mathbb{P}_{\vec{\sigma}_j}^{\otimes n}} \right] (d\omega) \\
& \quad \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^*) Q_0(dx) \mathbb{P}_{\vec{\sigma}}^{\otimes n}(d\omega), \tag{5.14}
\end{aligned}$$

where $\vec{\sigma}_{j,r} = (\sigma_1, \sigma_{j-1}, r, \sigma_{j+1}, \dots, \sigma_k)$ for $r \in \{0, 1\}$ and B_j is defined above. Now introduce binary valued functions:

$$\hat{f}(x) = \mathbf{1}(x \in \hat{G}_{n,m}) \text{ and } f_{\vec{\sigma}}^*(x) = \mathbf{1}(x \in G_K^*).$$

Then since $\sum_{l \neq j} \sigma_l \psi_l(x) \leq \psi_j(x) - \varphi_j(x)$ (see 5.2.2), we have coarsely for any $\vec{\sigma}$:

$$\forall x \in \chi_j, f_{\vec{\sigma}}^*(x) = \begin{cases} \sigma_j & \text{for } x \in B_j, \\ 0 & \text{otherwise,} \end{cases} \tag{5.15}$$

where $B_j = \{x \in \chi_j : \forall i |x_i - z_{j,i}| \leq \frac{\epsilon}{q}\}$. Now we go back to the lower bound. We can write:

$$\begin{aligned}
& \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^*) Q_0(dx) = \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(\hat{f} \neq f_{\vec{\sigma}}^*) Q_0(dx) \\
& \geq \mathbb{E}_{\nu(d\sigma_j)} \left[\int_{B_j} \mathbf{1}(\hat{f} \neq \sigma_j) Q_0(dx) \right] \\
& = \frac{1}{2} \left[\int_{B_j} [\mathbf{1}(\hat{f} \neq 1) + \mathbf{1}(\hat{f} \neq 0)] Q_0(dx) \right] \\
& = \frac{1}{2} \int_{B_j} Q_0(x) dx,
\end{aligned}$$

where we use (5.15) at the second line. Then it follows from (5.14) that:

$$\begin{aligned}
& \sup_{\vec{\sigma} \in \{-1, +1\}^k} \mathbb{E}_f \left\{ d_{\Delta}(\hat{G}_{n,m}, G_K^*) | Z_1^{(2)}, \dots, Z_m^{(2)} \right\} \\
& \geq \mathbb{E}_{\nu^{\otimes(k-1)}} \sum_{j=1}^k \int_{\Omega} \left[\frac{\mathbb{P}_{\vec{\sigma}_{j,0}}^{\otimes n}}{\mathbb{P}_{\vec{\sigma}_j}^{\otimes n}} \wedge \frac{\mathbb{P}_{\vec{\sigma}_{j,1}}^{\otimes n}}{\mathbb{P}_{\vec{\sigma}_j}^{\otimes n}} \right] (d\omega) \frac{1}{2} \int_{\chi_j} Q_0(dx) \mathbb{P}_{\vec{\sigma}}^{\otimes n}(d\omega) \\
& = \sum_{j=1}^k \mathbb{E}_{\nu^{\otimes(k-1)}} [1 - \mathbb{V}(\mathbb{P}_{\vec{\sigma},1}^{\otimes n}, \mathbb{P}_{\vec{\sigma},0}^{\otimes n})] \frac{1}{2} \int_{B_j} Q_0(dx) \\
& \geq \sum_{j=1}^k \mathbb{E}_{\nu^{\otimes(k-1)}} [1 - \sqrt{\chi^2(\mathbb{P}_{\vec{\sigma},1}^{\otimes n}, \mathbb{P}_{\vec{\sigma},0}^{\otimes n})}] \frac{1}{2} \int_{B_j} Q_0(dx) \\
& = \sum_{j=1}^k [1 - \sqrt{(1 + \chi^2(\mathbb{P}_1, \mathbb{P}_0))^n - 1}] \frac{1}{2} \int_{B_j} Q_0(dx), \tag{5.16}
\end{aligned}$$

where $\mathbb{P}_{\pm 1}$ is the law of $Z^{(1)}$ when $f = f_{\vec{\sigma}}$ with $\vec{\sigma} = (\pm 1, 1, \dots, 1)$. Then we can write, if $\chi^2(\mathbb{P}_1, \mathbb{P}_0) \leq \frac{C}{n}$:

$$\sup_{\sigma \in \{-1, +1\}^k} \mathbb{E}_{f_{\vec{\sigma}}, g_0} d_{\Delta}(\hat{G}_{n,m}, G_K^*) \geq c' \sum_{j=1}^k \int_{B_j} Q_0(dx) = c' kw, \quad (5.17)$$

where we use the definition of Q_0 .

Next step is to find a satisfying upper bound for $\chi^2(\mathbb{P}_1, \mathbb{P}_0)$. We have, by construction of $f_{\vec{\sigma}}$:

$$\begin{aligned} \chi^2(\mathbb{P}_1, \mathbb{P}_{-1}) &= \int \frac{[(f_{\vec{\sigma},+1} - f_{\vec{\sigma},-1})\mu * \eta]^2}{f_{\vec{\sigma},-1} * \eta} dx, \\ &\leq \int \frac{[(f_{\vec{\sigma},+1} - f_{\vec{\sigma},-1})\mu_0 * \eta]^2}{f_{\vec{\sigma},-1}\mu * \eta} dx + \int \frac{[(f_{\vec{\sigma},+1} - f_{\vec{\sigma},-1})\mu_1 * \eta]^2}{f_{\vec{\sigma},-1}\mu * \eta} dx. \end{aligned}$$

The r.h.s. term can be considered as engligible with a good choice of the parameters a and b . Hence, we concentrate on the first one. First remark that for all $x \in \mathbb{R}^d$

$$f_{\vec{\sigma},-1}\mu * \eta \geq C \frac{kw}{1+x^2}, \text{ and } \{(f_{\vec{\sigma},+1} - f_{\vec{\sigma},-1})\mu_0\} * \eta = q^{-\gamma} \{\psi_1 \mu_0\} * \eta(x) = q^{-\gamma} k\omega \{\psi_1 \rho\} * \eta(x).$$

Hence

$$\chi^2(\mathbb{P}_1, \mathbb{P}_{-1}) \leq Ckwq^{-2\gamma} \|(\psi_1 \rho) * \eta\|^2. \quad (5.18)$$

From the definition of ψ_1 and the conditions on η , we have in dimension $d = 2$ for simplicity:

$$\begin{aligned} \|(\psi_1 \rho) * \eta\|^2 &= \int (\psi_1 \rho) * \eta(x)^2 dx = \prod_{i=1}^2 \int |\mathcal{F}[\psi_1 \rho](t_i)|^2 |\mathcal{F}[\eta_i](t_i)|^2 dt_i \\ &= \prod_{i=1}^2 \int |\mathcal{F}[\rho(2\pi q \cdot)](t_i - 4\pi q)|^2 |\mathcal{F}[\eta_i](t_i)|^2 dt_i. \end{aligned}$$

Using (5.12), the noise assumption, and the fact that $q \rightarrow +\infty$, we get

$$\begin{aligned} &= Cq^{-2\bar{\beta}} \prod_{i=1}^2 \int |\mathcal{F}[\rho(2\pi q \cdot)](t_i - 4\pi q)|^2 dt_i, \\ &= Cq^{-2\bar{\beta}} \|\rho(2\pi q \cdot)\|^2, \\ &\leq Cq^{-2\bar{\beta}} \|\rho(2\pi q \cdot)\|^2 \leq Cq^{-2\bar{\beta}-2}. \end{aligned}$$

Using (5.18), one gets the following control of the quantity $\chi^2(\mathbb{P}_1, \mathbb{P}_{-1})$ in the general d -dimensional case:

$$\chi^2(\mathbb{P}_{\vec{\sigma},1}, \mathbb{P}_{\vec{\sigma},-1}) \leq Cq^{-2\gamma-\alpha\gamma-d-2\bar{\beta}} \leq \frac{C}{n}, \text{ with } q = n^{\frac{1}{2\gamma+\alpha\gamma+d+2\bar{\beta}}}. \quad (5.19)$$

Now using (5.17),

$$\sup_{\sigma \in \{-1, +1\}^k} \mathbb{E}_{f_{\vec{\sigma}}} d_{\Delta}(\hat{G}_{n,m}, G_K^*) \geq c'kw = c'q^{-\alpha\gamma} = c'n^{\frac{-\alpha\gamma}{2\gamma+\alpha\gamma+d+2\bar{\beta}}},$$

which concludes the proof of the second lower bound.

5.3 Proof of Theorem 3

The proof is presented for $d = 2$ for simplicity whereas straightforward modifications leads to the d -dimensional case. In the sequel, we identify each $\nu \in \Sigma'(\gamma_p, L)$ with a set $G_\nu = \{x : \nu(x) \geq 0\}$. By the same way, we identify G_K^* with $\nu^* = f - g$.

For all $G_\nu := \{\nu \geq 0\}$, we have, using the notations of Section 3:

$$\begin{aligned} & R_{n,m}(G_\nu) - R_{n,m}(G_K^*) - R_K^\lambda(G_\nu) + R_K^\lambda(G_K^*) \\ &= \frac{1}{2n} \sum_{i=1}^n U_i(G_\nu) + \frac{1}{2m} \sum_{i=1}^n V_i(G_\nu) := \frac{1}{2} T_n(G), \end{aligned}$$

where, for all $i \in \{1, \dots, n\}$,

$$U_i(G) = \{h_{G_K^*, \lambda}(Z_i^{(1)}) - h_{G_\nu, \lambda}(Z_i^{(1)})\} - \mathbb{E}[h_{G_K^*, \lambda}(Z_i^{(1)}) - h_{G_\nu, \lambda}(Z_i^{(1)})],$$

and

$$V_i(G) = \{h_{G_K^*, \lambda}(Z_i^{(2)}) - h_{G_\nu, \lambda}(Z_i^{(2)})\} - \mathbb{E}[h_{G_K^*, \lambda}(Z_i^{(2)}) - h_{G_\nu, \lambda}(Z_i^{(2)})].$$

Then, for all $i \in \{1, \dots, n\}$, using Lemmas 3 and 4 in Appendix we get

$$\mathbb{E}[U_i(G)]^2 \leq c \lambda_1^{-2\beta_1} \lambda_2^{-2\beta_2} d_\Delta(G, G_K^*) \leq c' \lambda_1^{-2\beta_1} \lambda_2^{-2\beta_2} d_{f,g}(G, G_K^*)^{\frac{\alpha}{\alpha+1}},$$

and

$$|U_i(G)| \leq C \prod_{i=1}^2 \lambda_i^{-\beta_i-1/2},$$

for some constant $C > 0$. The Bernstein's inequality leads to

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n U_i(G_\nu)\right| > a\right) \leq 2 \exp\left[-\frac{Cna^2}{a \times \lambda_1^{-\beta_1-1/2} \lambda_2^{-\beta_1-1/2} + \lambda_1^{-2\beta_1} \lambda_2^{-2\beta_2} d_\Delta(G_\nu, G_K^*)}\right],$$

for all $a > 0$. Since $\beta_i > 1/2$ for all $i \in \{1, \dots, d\}$, the particular choice $a = d_{f,g}(G_\nu, G_K^*)$ yields

$$\begin{aligned} P\left(\left|\frac{1}{n} \sum_{i=1}^n U_i(G_\nu)\right| > d_{f,g}(G, G_K^*)\right) &\leq 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} d_{f,g}(G, G_K^*)^{2-\frac{\alpha}{\alpha+1}}\right], \\ &= 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} d_{f,g}(G, G_K^*)^{\frac{2+\alpha}{\alpha+1}}\right]. \end{aligned}$$

Using the same algebra on the $V_i(G_\nu)$, we get

$$P(|T_n(G_\nu)| > d_{f,g}(G_\nu, G_K^*)) \leq 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} d_{f,g}(G, G_K^*)^{\frac{2+\alpha}{\alpha+1}}\right].$$

This concludes the first part of the proof. Let t a positive parameter which will be chosen further and introduce the set \mathcal{G}' defined as

$$\mathcal{G}' = \{G \in \mathcal{N}_{\delta_n}, d_{f,g}(G_K^*, G) > t \delta_n^{1+\alpha}\},$$

where \mathcal{N}_{δ_n} is the δ_n network introduced in Section 4.2. Using the upper bound above,

$$\begin{aligned} P\left(\exists G \in \mathcal{G}' : |T_n(G)| \geq \frac{1}{4} d_{f,g}(G, G_K^*)\right) &\leq \sum_{G \in \mathcal{G}'} P\left(|T_n(G)| \geq \frac{1}{4} d_{f,g}(G, G_K^*)\right), \\ &\leq \sum_{G \in \mathcal{G}'} 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} d_{f,g}(G, G_K^*)^{\frac{2+\alpha}{\alpha+1}}\right], \\ &\leq \sum_{G \in \mathcal{G}'} 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} (t \delta_n^{1+\alpha})^{\frac{2+\alpha}{\alpha+1}}\right], \\ &\leq \sum_{G \in \mathcal{G}'} 2 \exp\left[-Cn \lambda_1^{2\beta_1} \lambda_2^{2\beta_2} t^{\frac{2+\alpha}{\alpha+1}} \delta_n^{2+\alpha}\right]. \end{aligned}$$

Since $\log \text{card}(\mathcal{N}_{\delta_n}) = A\delta_n^{-2/\gamma}$, we get

$$P\left(\exists G \in \mathcal{G}' : |T_n(G)| \geq \frac{1}{4}d_{f,g}(G, G_K^*)\right) \leq \exp\left[A\delta_n^{-2/\gamma} - Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2}t^{\frac{2+\alpha}{\alpha+1}}\delta_n^{2+\alpha}\right].$$

Thanks to the value of δ_n ,

$$\delta_n^{-2/\gamma} \sim n\lambda_1^{2\beta_1}\lambda_2^{2\beta_2}\delta_n^{2+\alpha}.$$

Hence

$$P\left(\exists G \in \mathcal{G}' : |T_n(G)| \geq \frac{1}{4}d_{f,g}(G, G_K^*)\right) \leq \exp\left[-C\delta_n^{-2/\gamma}\right] = \exp\left[-C\left(\frac{\lambda_1^{-\beta_1}\lambda_2^{-\beta_2}}{\sqrt{n}}\right)^{-2/\gamma \cdot \frac{2}{2/\gamma+2+\alpha}}\right].$$

Now, using Lemma 2 in Appendix, we can find a set $G_n \in \mathcal{N}_{\delta_n}$ such that:

$$d_{f,g}(G_K^*, G_n) \leq \|\nu^* - \nu_n\|_\infty^{\alpha+1} \leq C_0\delta_n^{1+\alpha},$$

for some positive constant C_0 . Then, for all $G \in \mathcal{G}'$, we get

$$\frac{1}{4}d_{f,g}(G, G_K^*) - \frac{1}{2}d_{f,g}(G_n, G_K^*) \geq \frac{t}{4}\delta_n^{1+\alpha} - \frac{C}{2}\delta_n^{1+\alpha} \geq \frac{C}{2}\delta_n^{1+\alpha}, \quad (5.20)$$

provided that $t > 4C_0$. We eventually obtain:

$$\begin{aligned} & P\left(d_{f,g}(G_K^*, \hat{G}_n) > t\delta_n^{1+\alpha}\right) \\ & \leq P\left(\exists G \in \mathcal{G}' \text{ such that } R_n(G) \leq R_n(G_n)\right), \\ & = P\left(\exists G \in \mathcal{G}' \text{ such that } \frac{1}{2}d_{f,g}^\lambda(G, G_K^*) + Z_n(G) - \frac{1}{2}d_{f,g}^\lambda(G_n, G_K^*) - Z_n(G_n) \leq 0\right) \end{aligned} \quad (5.21)$$

where for all $G_1, G_2 \subset K$, $d_{f,g}^\lambda(G_1, G_2)$ is defined as

$$\frac{1}{2}d_{f,g}^\lambda(G_1, G_2) = R_K^\lambda(G_1) - R_K^\lambda(G_2).$$

Last step is to control the bias term. For all $G_1, G_2 \subset K$, we can remark that

$$\begin{aligned} & \left|(R_K^\lambda - R_K)(G_1 - G_2)\right| \\ & \leq \left|\int \left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{z-x}{\lambda}\right)f(z)dQ(z) - f(x)\right] [\mathbf{1}(x \in G_1^C) - \mathbf{1}(x \in G_2^C)] dQ(x) \right. \\ & \quad \left. + \int \left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{z-x}{\lambda}\right)g(z)dQ(z) - g(x)\right] [\mathbf{1}(x \in G_1) - \mathbf{1}(x \in G_2)] dQ(x) \right| \\ & \leq \int_{G_1 \Delta G_2} |\mathcal{K}_\lambda * \nu(x) - nu(x)| dQ(x), \\ & \leq \|\mathcal{K}_\lambda * \nu - \nu\|_\infty d_\Delta(G_1, G_2), \\ & \leq Cd_\Delta(G_1, G_2) [\lambda_1^\gamma + \lambda_2^\gamma], \end{aligned}$$

for some $C > 0$, provided that for $\nu \in \Sigma(\gamma, L)$ and \mathcal{K} a kernel of order $l = \lfloor \gamma \rfloor$:

$$\|\mathcal{K}_\lambda * \nu - \nu\|_\infty \leq C [\lambda_1^\gamma + \lambda_2^\gamma]. \quad (5.22)$$

Using the Young inequality

$$xy^r \leq ry + (1-r)x^{1/(1-r)},$$

with $r = \alpha/(\alpha+1)$, we get for all $G_1, G_2 \subset K$

$$\left|(R_K^\lambda - R_K)(G_1 - G_2)\right| \leq (1-r)\gamma^{1/(1-r)} [\lambda_1^2 + \lambda_2^2]^{\frac{\gamma(1+\alpha)}{2}} + \gamma^{-1/r}d_{f,g}(G_1, G_2), \quad (5.23)$$

for some $\gamma > 0$. Hence, it follows from (5.20)-(5.21) that

$$\begin{aligned}
& P\left(d_{f,g}(G_K^*, \hat{G}_n) > t\delta_n^{1+\alpha}\right) \\
& \leq P\left(\exists G \in \mathcal{G}' \text{ such that } \frac{3}{4}d_{f,g}(G, G_K^*) + Z_n(G) - \frac{1}{4}d_{f,g}(G_n, G_K^*) - Z_n(G_n) + C \sum_{i=1}^2 \lambda_i^{\nu(1+\alpha)} \leq 0\right), \\
& \leq P\left(\exists G \in \mathcal{G}' \text{ such that } Z_n(G) \leq -\frac{1}{4}d_{f,g}(G, G_K^*)\right) + P\left(Z_n(G_n) \geq C(\delta_n^{1+\alpha} + \sum_{i=1}^2 \lambda_i^{\nu(1+\alpha)})\right).
\end{aligned}$$

In order to conclude, remark that the proposed choice of $(\lambda_j)_{j=1\dots d}$ provides

$$\delta_n^{1+\alpha} \simeq \sum_{i=1}^2 \lambda_i^{\gamma(1+\alpha)} \Leftrightarrow \forall i \in \{1, 2\}, \left(\frac{1}{\lambda_i^\beta \sqrt{n}}\right)^{\frac{2}{2/\nu+2+\alpha}} \simeq [\lambda_1^2 + \lambda_2^2]^{\gamma/2}.$$

The end of the proof follows exactly the same lines as [2].

5.4 Proof of Theorem 4

Let us prove the first assertion. Using the definition of \hat{G}_n^λ in (1.7), we have:

$$\begin{aligned}
d_{f,g}(\hat{G}_n^\lambda, G_K^*) & \leq R_K(\hat{G}_n^\lambda) - R_K(G_K^*) - R_n^\lambda(\hat{G}_n^\lambda) + R_n^\lambda(G_K^*) \\
& \leq R_K^\lambda(\hat{G}_n^\lambda) - R_n^\lambda(\hat{G}_n^\lambda) + R_n^\lambda(G_K^*) - R^\lambda(G_K^*) + (R_K - R_K^\lambda)(\hat{G}_n^\lambda - G_K^*).
\end{aligned}$$

Consider the empirical processes $\nu_n^{(j)}$, for $j \in \{1, 2\}$, defined as:

$$\nu_n^{(j)}(G) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_{G,\lambda}(Z_i^{(j)}) - \mathbb{E}h_{G,\lambda}(Z^{(j)}). \quad (5.24)$$

Hence we can write:

$$\begin{aligned}
& \int (f - g)(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}) \\
& \leq \frac{1}{2\sqrt{n}}(\nu_n^{(1)}(G^{*C}) - \nu_n^{(1)}(\hat{G}_n^{\lambda C})) + \frac{1}{2\sqrt{n}}(\nu_n^{(2)}(G^*) - \nu_n^{(2)}(\hat{G}_n^\lambda)).
\end{aligned}$$

Now denoting $\Lambda = \Pi_{i=1}^d \lambda_i^{-\beta_i - \frac{1}{2}}$, $c(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i}$ and $\rho = 2/\gamma$, consider the event:

$$\Omega = \{d_\Delta(\hat{G}_n, G_K^*) \geq c(\lambda)^{-\frac{2}{1+\rho}} n^{-\frac{1}{1+\rho}} \Lambda^{\frac{2}{1+\rho}}\}.$$

We have on Ω , using both the margin assumption and Lemma 3:

$$\begin{aligned}
& \int (f - g)(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}) \\
& \leq \frac{d_{\Delta}^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^*)c(\lambda)}{\sqrt{n}} \left[\frac{\nu_n^{(1)}(G^{*C}) - \nu_n^{(1)}(\hat{G}_{n,m}^{\lambda C})}{c(\lambda)d_{\Delta}^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^*) \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}} \right. \\
& \quad \left. + \frac{\nu_n^{(2)}(G^*) - \nu_n^{(2)}(\hat{G}_{n,m}^\lambda)}{c(\lambda)d_{\Delta}^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^*) \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}} \right] \\
& \leq \frac{d_{f,g}^{\frac{1-\rho}{2} \frac{\alpha}{\alpha+1}}(\hat{G}_{n,m}^\lambda, G^*)c(\lambda)}{\sqrt{n}} [V_n^{(1)} + V_n^{(2)}],
\end{aligned}$$

where $V_n^{(j)}$ is the random variable defined as, for $j \in \{1, 2\}$:

$$V_n^{(j)} = \sup_{G \in \mathcal{G}} \frac{|\nu_n^{(j)}(G^*) - \nu_n^{(j)}(G)|}{c(\lambda)^\rho \|h_G^\lambda - h_{G^*}^\lambda\|_{2, Z^{(j)}}^{1-\rho} \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}}. \quad (5.25)$$

A generalization of Lemma 5.13 of [16] provides that the variable $V_n^{(1)} + V_n^{(2)}$ has controled moments. We can write, using Young's inequality $xy^r \leq ry + (1-r)x^{1/(1-r)}$ for $r = \frac{1-\rho}{2} \frac{\alpha}{\alpha+1}$:

$$\int (f - g)(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}}) \leq c \left(\frac{c(\lambda)}{\sqrt{n}} [V_n^{(1)} + V_n^{(2)}] \right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}}. \quad (5.26)$$

Finally we conclude that on Ω :

$$d_{f,g}(\hat{G}_n^\lambda, G_K^*) \leq c \left(\frac{c(\lambda)}{\sqrt{n}} [V_n^{(1)} + V_n^{(2)}] \right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + \sup_{G \in \mathcal{G}(\gamma, L)} |R_K - R_K^\lambda|(G).$$

This allows us to get the first assertion of the Theorem since we have on Ω^C , we have from an easy calculation:

$$d_\Delta(\hat{G}, G_K^*) \leq c(\lambda)^{-\frac{2}{1+\rho}} n^{-\frac{1}{1+\rho}} \Lambda^{\frac{2}{1+\rho}} = c(\lambda)^{\frac{1}{1-\rho}} n^{-\frac{1}{1-\rho}} \leq \left(\frac{c(\lambda)}{\sqrt{n}} \right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}}$$

provided that $\lambda = (\lambda_1, \dots, \lambda_d)$ is choosen small enough to ensure the last inequality.

To get the second assertion, we apply Lemma 1 with $G_1 = \hat{G}_{n,m} := \hat{G}$ and $G_2 = G_K^*$ and Lemma 4 in order to get

$$\begin{aligned} d_\Delta(\hat{G}_{n,m}, G_K^*) &\leq C \|\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_{n,m}\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}}\|_1 + \lambda_1^\gamma + \lambda_2, \\ &\leq C \left(\int |f - g| \left| \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} \right| \right)^{\alpha/\alpha+1} + \lambda_1^\gamma + \lambda_2. \end{aligned} \quad (5.27)$$

Since $\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G\}} \in [0, 1]$ for any $G \in \mathcal{G}(\gamma, L)$, we get

$$\begin{aligned} &\int |f - g| |\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}| \\ &\leq \int |f - g| |\mathbf{1}_{G_K^*} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}| + \int |f - g| |\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathbf{1}_{G_K^*}| \\ &= \int (f - g)(\mathbf{1}_{G_K^*} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}) + \int |f - g| |\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathbf{1}_{G_K^*}|, \\ &\leq \int (f - g) (\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in \hat{G}_n\}}) + 2 \int |f - g| |\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}} - \mathbf{1}_{G_K^*}|. \end{aligned}$$

which gives, gathering with (5.27):

$$\begin{aligned} &d_\Delta(\hat{G}_{n,m}, G_K^*) \\ &\leq C \left(R_K^\lambda(\hat{G}_n^\lambda) - R_K^\lambda(G_K^*) + 2 \int |f - g| |\mathbf{1}_{G_K^*} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_K^*\}}| \right)^{\alpha/\alpha+1} + C(\lambda_1^\gamma + \lambda_2). \end{aligned} \quad (5.28)$$

Finally using Lemma 1, (5.28) and (5.26), we have on Ω :

$$d_\Delta(\hat{G}, G_K^*) \leq c \left(\frac{c(\lambda)}{\sqrt{n}} [V_n^{(1)} + V_n^{(2)}] \right)^{\frac{2\alpha}{\alpha+2+\rho\alpha}} + \sum_{i=1}^{d-1} \lambda_i^\gamma + \lambda_d + \left(\int |f - g| |\mathbf{1}_{G_K^*} - h_{G_K^*}| \right)^{\alpha/\alpha+1}.$$

Integrating the above inequality, we conclude the proof noting that on Ω^C , we have from an easy calculation:

$$d_\Delta(\hat{G}, G_K^*) \leq c(\lambda)^{-\frac{2}{1+\rho}} n^{-\frac{1}{1+\rho}} \Lambda^{\frac{2}{1+\rho}} = \left(\prod_{i=1}^d \lambda_i n \right)^{\frac{-1}{1-\rho}} \leq n^{-\tau_d(\alpha, \beta, \gamma)},$$

provided that $2\beta_1 + 2\beta_2\gamma + 1 \geq \gamma$, or in particular when $\beta_2 \geq \frac{1}{2}$.

6 Appendix

For the sake of convenience, we assume throughout this section that the kernels $(\mathcal{K}_j)_{j=1\dots d}$ are compactly supported. This assumption can easily be relaxed up to a more complicated algebra.

Lemma 1 *For any $G_1, G_2 \in \mathcal{G}(\gamma, L)$, and any $\lambda > 0$, we have:*

$$d_\Delta(G_1, G_2) \leq c \|\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}\|_1 + \left(\sum_{i=1}^{d-1} \lambda_i^2 \right)^{\gamma/2} + \lambda_d.$$

PROOF. For the sake of convenience, we only give the proof in the particular case where $d = 2$. Using the equality $|a - b| = a + b - 2 \min(a, b)$, $\forall a, b \in \mathbb{R}$, we can write:

$$\begin{aligned} & \|\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}\|_1 \\ &= \int \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}} + \int \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}} - 2 \int \min(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}}, \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}). \end{aligned}$$

Then for any $G \in \mathcal{G}(\gamma, L)$, remark that

$$\begin{aligned} & \left| \int_{[0,1]} \mathcal{K}_\lambda * \mathbf{1}_{\{x \in G\}} dx_2 - b(x_1) \right| \\ &= \left| \int_{[0,1]} \int_{\mathbb{R}^2} \frac{1}{\lambda} \mathcal{K} \left(\frac{u-x}{\lambda} \right) \mathbf{1}(u_2 \leq b(u_1)) du dx_2 - b(x_1) \right|, \\ &\leq \left| \int_{[0,1]} \int_{\mathbb{R}} \frac{1}{\lambda_1} \mathcal{K} \left(\frac{u_1 - x_1}{\lambda_1} \right) \mathbf{1}(x_2 \leq b(u_1)) du_1 dx_2 - b(x_1) \right| + C\lambda_2, \\ &\leq \left| \int_{\mathbb{R}} \frac{1}{\lambda_1} \mathcal{K} \left(\frac{u_1 - x_1}{\lambda_1} \right) b(u_1) du_1 - b(x_1) \right| + C\lambda_2, \\ &\leq C(\lambda_1^\gamma + \lambda_2). \end{aligned} \tag{6.1}$$

Moreover, noticing that $\int \min(f, g) \leq \min(\int f, \int g)$, we have, using (6.1)

$$\begin{aligned} & \int \min(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}}, \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}) \\ &\leq \int_{[0,1]} \min \left(\int_{[0,1]} \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}} dx_2, \int_{[0,1]} \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}} dx_2 \right) dx_1, \\ &\leq \int_0^1 \min(b_1(x_1), b_2(x_1)) dx_1 + C(\lambda_1^\gamma + \lambda_2). \end{aligned}$$

Finally we arrive at the conclusion:

$$\begin{aligned} d_\Delta(G_1, G_2) &= \int |b_1 - b_2| = \int b_1 + \int b_2 - 2 \int \min(b_1, b_2) \\ &\leq \int b_1 + \int b_2 - 2 \int \min(\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}}, \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}) + 2C[\lambda_1^\gamma + \lambda_2] \\ &\leq \|\mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_1\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{\cdot \in G_2\}}\|_1 + 2C[\lambda_1^\gamma + \lambda_2], \end{aligned}$$

for some positive constant C . □

Lemma 2 *For any (f, g) satisfying the margin assumption with parameter $\alpha > 0$, we have:*

$$d_{f,g}(G_\nu, G_K^*) \leq \|\nu - \nu^*\|_\infty^{\alpha+1},$$

where $G_\nu = \{\nu \geq 0\}$ and $\nu^* = f - g$.

The proof is a straightforward modification of the proof of Lemma 5.1 in [2] which state a similar result in the binary classification framework.

Lemma 3 *Assume that η satisfies the Noise assumption and let \mathcal{K}_η the deconvolution kernel. Then we have,*

$$(i) \quad \mathbb{E}[h_{G,\lambda}(Z) - h_{G',\lambda}(Z)]^2 \leq d_\Delta(G, G') \prod_{i=1}^d \lambda_i^{-2\beta_i}.$$

$$(ii) \quad \sup_{x \in K} |h_{G,\lambda}(x) - h_{G',\lambda}(x)| \leq \prod_{i=1}^d \lambda_i^{-\beta_i - 1/2}$$

PROOF For the sake of convenience, we only consider the case where $d = 1$. We first prove (i). We have:

$$\begin{aligned} \mathbb{E}[h_{G,\lambda}(Z) - h_{G',\lambda}(Z)]^2 &= \int_K \left[\int_{\mathbb{R}} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) (\mathbf{1}_{\{x \in G\}} - \mathbf{1}_{\{x \in G'\}}) \mathbf{1}_{\{x \in K\}} dQ(x) \right]^2 (f\mu) * \eta(z) dz, \\ &\leq c \int_{\mathbb{R}} \frac{1}{\lambda^2} |\mathcal{F}[\mathcal{K}_\eta(\cdot/\lambda)](t)|^2 |\mathcal{F}[\mu \times (\mathbf{1}_{\{\cdot \in G\}} - \mathbf{1}_{\{\cdot \in G'\}}) \mathbf{1}_{\{\cdot \in K\}}](t)|^2 dt, \\ &\leq C \lambda^{-2\beta} \int_K |\mu(t)|^2 \mathbf{1}_{\{t \in G \Delta G'\}} dt, \\ &\leq C \lambda^{-2\beta} d_\Delta(G, G'). \end{aligned}$$

Indeed, for all $s \in \mathbb{R}$

$$\frac{1}{\lambda^2} |\mathcal{F}[\mathcal{K}_\eta(\cdot/\lambda)](s)|^2 = |\mathcal{F}[\mathcal{K}_\eta](s\lambda)|^2 \leq \sup_{t \in \mathbb{R}} \left| \frac{\mathcal{F}[\mathcal{K}](t\lambda)}{\mathcal{F}[\eta](t)} \right|^2 \leq \sup_{t \in [-\frac{1}{\lambda}; \frac{1}{\lambda}]} \left| \frac{1}{\mathcal{F}[\eta](t)} \right|^2 \leq C \lambda^{-2\beta}, \quad (6.2)$$

provided that \mathcal{K} has compact Fourier transform. By the same way,

$$\begin{aligned} \sup_{x \in \mathbb{R}} |h_{G,\lambda}(x) - h_{G',\lambda}(x)| &= \sup_{x \in \mathbb{R}} \int_{G \Delta G'} \frac{1}{\lambda} \left| \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \right| dQ(x), \\ &\leq \sup_{x \in \mathbb{R}} \int_K \frac{1}{\lambda} \left| \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \right| dx, \\ &\leq C \sup_{x \in \mathbb{R}} \sqrt{\int \frac{1}{\lambda^2} \mathcal{K}_\eta^2 \left(\frac{z-x}{\lambda} \right) dx} \leq \lambda^{-\beta-1/2}, \end{aligned}$$

where the last line is inspired by (6.2). □

The following Lemma proposes a generalization to the well-known inequality of [24].

Lemma 4 *Let h a positive and bounded function integrable with respect to Q with $\|h\|_\infty \leq B$. Suppose the margin assumption holds and denote by $\alpha > 0$ the margin parameter. Then, there exists positive constants $c(\alpha)$ and $C(\alpha)$ such that:*

$$c(\alpha) \left(\int h(x) dQ(x) \right)^{\frac{\alpha+1}{\alpha}} \leq \int |f-g| h(x) dQ(x) \leq C(\alpha) \int h(x) dQ(x).$$

In particular, for all $G_1, G_2 \subset K$, we have

$$c(\alpha) (d_\Delta(G_1, G_2))^{\frac{\alpha+1}{\alpha}} \leq d_{f,g}(G_1, G_2) \leq C(\alpha) d_\Delta(G_1, G_2).$$

PROOF. The proof follows exactly the proof of Lemma 2 of [24]. Since $Q(K)$ is bounded, $Q(|f - g| \leq \eta) \leq c_2 \eta^\alpha$ for $0 < \eta < \eta_0$ implies $Q(|f - g| < \eta) \leq \tilde{c}_2 \eta^\alpha, \forall \eta > 0$ where $\tilde{c}_2 := \tilde{c}_2(\alpha, c_2, \eta_0, Q(K))$. Then we have since $h > 0$ is bounded, choosing $\eta = \left(\frac{\int h}{2B\tilde{c}_2}\right)^{\frac{1}{\alpha}}$:

$$\begin{aligned}
\int |f - g| h dQ &\geq \int |f - g| \mathbf{1}(|f - g| \geq \eta) h dQ \\
&\geq \eta \left(\int h dQ - \int h \mathbf{1}(|f - g| < \eta) dQ \right) \\
&\geq \eta \left(\int h dQ - BQ(|f - g| < \eta) \right) \\
&\geq \eta \left(\int h dQ - \tilde{c}_2 B \eta^\alpha \right) \\
&= c(\alpha) \left(\int h dQ \right)^{\frac{\alpha+1}{\alpha}},
\end{aligned}$$

where $c(\alpha) = 2^{-1-1/\alpha} (B\tilde{c}_2)^{-1/\alpha}$. The upper bound is straightforward since $|f - g|$ is bounded from above. □

References

- [1] J-Y. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris VI and VII., 2004.
- [2] J-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35:608–633, 2007.
- [3] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.
- [4] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.
- [5] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101 (473):138–156, 2006.
- [6] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.
- [7] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- [8] C. Butucea. goodness-of-fit testing and quadratic functional estimation from indirect observations. *Annals of Statistics*, 35:1907–1930, 2007.
- [9] Olivier Chapelle, Jason Weston, Léon Bottou, L Eon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, pages 416–422. MIT Press, 2001.
- [10] A. Delaigle and I. Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, 56:19–47, 2004.

- [11] A. Delaigle and I. Gijbels. Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, 16:773–788, 2006.
- [12] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
- [13] W.H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 2000.
- [14] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272, 1991.
- [15] J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *Annals of Statistics*, 21 (4):1900–1925, 1993.
- [16] S. Van De Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.
- [17] Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry A. Wasserman. Minimax manifold estimation. *CoRR*, abs/1007.0549, 2010.
- [18] J. Klemela and E. Mammen. Empirical risk minimization in inverse problems. *Annals of Statistics*, 38 (1):482–511, 2010.
- [19] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34 (6):2593–2656, 2006.
- [20] A.P. Korostelev and A.B. Tsybakov. *Minimax theory of Image Reconstruction. Lecture Notes in Statistics*. Springer Verlag, 1993.
- [21] B. Laurent, J.M. Loubes, and C. Marteau. Testing inverse problems: a direct or an indirect problem? *Journal of Statistical Planning and Inference*, 141:1849–1861, 2011.
- [22] J.M. Loubes and C. Marteau. Goodness-of-fit strategies from indirect observations. *Working paper*.
- [23] S. Loustau. Penalized empirical risk minimization over besov spaces. *Electronic journal of Statistics*, 3:824–850, 2009.
- [24] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.
- [25] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006.
- [26] A. Meister. *Deconvolution problems in nonparametric statistics*. Springer-Verlag, 2009.
- [27] S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.
- [28] Y. Tang. Minimax nonparametric classification - part i: Rates of convergence, part ii: Model selection for adaptation. *IEEE Trans. Inf. Theory*, 45:2271–2292, 1999.
- [29] A.B. Tsybakov. *Introduction à l’estimation non-paramétrique*. Springer-Verlag, 2004.
- [30] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004.
- [31] A.B. Tsybakov and S.A. Van De Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33 (3):1203–1224, 2005.

- [32] A. W. van der Vaart and J. A. Weelner. *Weak convergence and Empirical Processes. With Applications to Statistics*. Springer Verlag, 1996.
- [33] V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science, Springer, 2000.